2017

Urban flood modelling using geo-social intelligence

Kun Yang
University of Wollongong

Urban Flood Modelling Using Geo-social Intelligence

A thesis submitted in partial fulfilment

for the award of the degree Master of

Philosophy

from

University of Wollongong

by

KUN YANG

4814265

School of Computing and Information Technology

March 2017

Certification

I, Kun Yang, declare that this thesis, submitted in partial fulfilment of the requirements for the award of Master of Philosophy (Information Technology), in the School of Computing and Information Technology, University of Wollongong, is wholly my own work unless otherwise referenced or acknowledged. The document has not been submitted for qualifications at any other academic institution.

Kun Yang

28 March 2017

Acknowledgements

I would like to thank the University of Wollongong's SMART Infrastructure Facility for their ongoing support with my Masters of Philosophy research. I could not have hoped to have worked on a more interesting project that had real world application, than PetaJakarta.org. PetaJakarta was a UOW defined project, that was awarded a competitive Twitter Data Grant that meant my work had a real impact in Jakarta, Indonesia, helping local authorities to respond to seasonal monsoonal flooding, and raising awareness among the community on how geo-social intelligence could be used to harness collective intelligence.

While at the University of Wollongong, I was situated in the SMART Infrastructure Facility with direct access to Dr Tomas Holderness, a chartered geographer and Vice Chancellor Fellow from whom I am indebted to for his patience and vision. Tomas Holderness together with the co-director of PetaJakarta Dr Etienne Turpin, is now at the MIT Urban Risk Lab. Tomas was responsible for the three specific experiments that I had to complete as part of this MPhil. He led my technical contributions into the wider PetaJakarta project, and quite unbelievably I saw the fruits of my work being implement live into Jakarta between June 2013 and June 2016. Sometimes timing is everything, and I feel I hit the jackpot.

To my primary supervisor, Professor Katina Michael who patiently took me through the research process, explained the fundamentals of GIS and database principles in the context of social media, and crowdsourcing urban data, a very big thank you! You were there for me every step of the way, inspiring me toward completion. I miss our weekly meetings. Katina you helped me realise my dreams, and you never let me lose sight of the bigger picture. You are the reason why I am now back in China thriving in my own organisation with angel investors backing me up. You listened so carefully to all my thoughts, ideas, beyond the thesis. I will never ever forget what you have done for me. Associate Professor Jun Shen, thank you for your final year's co-supervision when Dr Holderness left for MIT. I enjoyed completing coursework with you and your ongoing advice about how to prepare for the final write up. I also appreciate the assistance of Dr Roba Abbas who edited the last draft of my work. And of course the lecturers who taught me in 30 credit points I completed toward this degree.

Finally, to my mother and father, thank you. Without your support, none of this would be possible. You helped me every day of my life. I dedicate this work to you.

Online Contributions

The PetaJakarta Project, though meant to solve an urban flooding problem, has had a global reach. It has changed the way that government and non-government organisations approach disasters in megacities by raising the possibility of using social media to respond and then forward plan for future events. The Jakarta Emergency Management Agency (BPBD DKI Jakarta) was a direct client of my work, although my segment was a small part of the project at large, it was an integral piece into analysing the relationship between Tweets and actual flood zones. Please see the following reputable media publications of which my research directly contributed. Rapid application development meant rapid deployment was possible.

An online tool for civic engagement during emergency events: http://smart.uow.edu.au/projects/petajakarta-org/index.html

Social data mapped in real time helping to save lives and inform emergency services during extreme weather events: https://www.uow.edu.au/research/researchimpact/UOW208258.html

Harnessing social media to respond to natural disasters: https://globalchallenges.uow.edu.au/impact/UOW219619.html

Peta Jakarta: real-time flood mapping in Jakarta: http://theodi.org/odi-showcase-peta-jakarta-real-time-flood-mapping-jakarta

PetaJakarta: what works: http://www.newcitiesfoundation.org/global-urban-innovators-alumni/petajakarta/

PetaJakarta case study: https://aws.amazon.com/solutions/case-studies/petajakarta/

How tweeting about floods became a civic duty in Jakarta: https://www.theguardian.com/public-leaders-network/2016/jan/25/floods-jakarta-indonesia-twitter-petajakarta-org

Need the latest news on flooding? In Jakarta, there's an app for that: https://www.pri.org/stories/2016-09-16/need-latest-news-flooding-jakarta-theres-app

Peta Jakarta gets netizens to report floods: http://www.thejakartapost.com/news/2015/02/04/peta-jakarta-gets-netizens-report-floods.html

Jakarta: A City on the Edge of a Social Media Revolution: http://voices.nationalgeographic.com/2016/02/10/jakarta-a-city-on-the-edge-of-a-social-media-revolution/

Dari Jakarta sampai London: Bagaimana kota-kota dunia melawan polusi udara? http://www.bbc.com/indonesia/majalah-39176908

Antara doeloe: peta Jakarta versi swasta banyak keliru: http://www.antaranews.com/berita/587014/antara-doeloe-peta-jakarta-versi-swasta-banyak-keliru

Making smart cities work for people. No 1: Crowdsourcing flood maps in Jakarta: http://www.citymetric.com/horizons/making-smart-cities-work-people-no-1-crowdsourcing-flood-maps-jakarta-1228

This site lets you know when Jakarta's streets turn to rivers: https://www.techinasia.com/indonesia-jakarta-floods-peta-app

Be a Global Urban Innovator! A Call for the World's Best Ideas to Make our Cities Better: http://www.huffingtonpost.com/adam-cutts/be-a-global-urban-innovat_b_8837736.html

Banjir Jakarta Bisa Dipantau dari PetaJakarta.org: http://www.cnnindonesia.com/teknologi/20141202180741-185-15349/banjir-jakarta-bisa-dipantau-dari-petajakartaorg/

Corporate Social Responsibility for a Data Age: https://ssir.org/articles/entry/corporate_social_responsibility_for_a_data_age

Ini peta banjir di Jakarta berdasarkan aduan masyarakat: http://www.antaranews.com/berita/479026/ini-peta-banjir-di-jakarta-berdasarkan-aduan-masyarakat

PetaJakarta turns tweets into flood alerts – ANDS: http://www.ands.org.au/news-and-events/dataimpact/data-impact-stories/petajakarta-turns-tweets-into-flood-alerts

Additionally, I have submitted a full research paper for consideration to the forthcoming *International Symposium on Technology and Society 2017*, to be held in Sydney, Australia in August 2017.

Abstract

Social media is not only a way to share information among a group of people but also an emerging source of rich primary data that can be crowdsourced for good. The primary function of social media is to allow people to network near real-time, yet the repository of amassed data can also be applied to decision support systems in response to extreme weather events.

The megacity of Jakarta, Indonesia has the greatest number of Twitter users of any city worldwide. Furthermore, the city experiences seasonal flooding during the annual monsoon seasons, which endangers human health, damages civic infrastructure and causes large economic loss. In this study, the use of social media data is examined as a way to help government organisations respond to flooding in a timely manner.

Tweets from two previous monsoons related to flooding were collected and analysed using the hashtag (#) “banjir”. Additionally, government data sources on the location of flood events in the city over the same period were collected. By analysing the relationship between the tweets and the flood events, this study aims to create “trigger metrics” of flooding based on Twitter activity. Such trigger metrics have the advantage of being able to provide a situational overview of flood conditions in near real-time, as opposed to formal government flood maps which are only produced on a six-hourly schedule. The aim is to provide continuous intelligence, rather than discrete intervals of decision-making capability.

The theory of this thesis demonstrated is to enhance the capacity to understand and promote the resilience of cities to both extreme weather events due to climate change and to long-term infrastructure transformation with the process of climate adaptation. To understand the full potential of Twitter as a real-time indicator for flooding, this research aims to quantify the temporal relationship between tweets related to flooding, and flood events in the city of Jakarta, Indonesia. This research can also provide methodological support to make social media a real time crowd-sourcing tool during extreme weather events. Past Twitter data and the real observed flooding events are used as the basis for modelling the urban flooding event in its totality.

To advance the comprehensive understanding of the relationship between flood events and Twitter activity in Jakarta and the modelling of urban flood using geo-social intelligence, this

research will quantify the relationship using three stages. Each stage is characterised by an experiment, the results of which are presented and interpreted. This thesis presents the background of the research topic, literature review, research methodology, process of analysis and statistical analysis of results based on these experiments, and a discussion of the outcomes of the research.

Chapter 1. Introduction

1.1 Definitions

This study is located in the domain of geographic information systems (GIS). GIS use "information technology and data to input, structure, manipulate, integrate, analyse, and display information with a geospatial aspect"[1]. In this case, the University of Wollongong PetaJakarta Global Challenge Grant provides the project scope, bringing together multiple rich data sets from several sources, such as social media (i.e. Twitter data), and government reports (e.g. flood detail data). GIS has the primary ability to bring together information and join it to digital imagery which enables spatial analysis. The data in a GIS that is "tagged" to a location is known as spatiotemporal data because it includes both time and space information.

Take for example, a standard tweet which contains text and optionally one or more hashtags or identifiers within the body of the message. It also contains a time stamp, a date stamp, a location stamp, and possibly a picture or multimedia clip. Geographically locked to a position on the earth's surface, data can reveal a great deal about context, and even physical changes and human movement changes in short intervals of space and time. A variety of analyses can then be performed on the data, taking the form of traditional statistical analysis (like in any other information system), or spatial analysis that takes advantage of the geographic component embedded in the system. In spatial analysis, **"**knowledge of a process is used to predict the spatial patterns that might result, and the likelihood of any observed pattern being a result of this process is then established by an analysis of one or more of its realizations" [2].

The significance of this study within the context of the PetaJakarta project is in the use of social media for civic infrastructure management. In this case, it is the use of Twitter data to aid in the improvement of civic infrastructure management for the provision of public services for citizens and the protection of the city from extreme weather events. While there are a variety of social media tools and platforms that give people the ability to create, share or exchange information, text messages, pictures and videos in virtual communities and networks, Twitter (and tweets) belong to a set of computer-mediated applications known as micro-blogging tools.

These applications allow “users to send short messages to people subscribed to their streams”[3]. Microblogs allow for very succinct text messages to be sent, usually 140 characters in length. Increasingly, citizens are participating and generating content that can be used for analysis, and not all of it is text-based. This is known as Volunteer Geographic Information (VGI) and it is the harnessing of applications and tools to create, assemble and disseminate geographic data that is provided voluntarily by individuals in the community freely [4]. In some ways it can be considered an information flow from the people to the people, where there is stealth in the collection of data analysed in an aggregated fashion.

Tomas Holderness describes the outcome of VGI data as geosocial intelligence to aid in decision-making [5]. Increasingly, given the sensor data available on handsets, more and more people are enabling location capabilities, for instance, when tweeting photos or even sending a plain message. These kinds of networks that are increasingly become dependent on the location element are known as location-based social networks (LBSN). For further reading on LBSN, including their social and ethical implications, refer to the work of Fusco et al. [6] and [7].

1.2 Research Background

Flooding in Jakarta is a problem that endangers human health, damages civic infrastructure and causes substantial economic loss. Social media companies like Twitter are now seeking to harness their data to solve big problems, for instance in the disaster management field. The University of Wollongong was awarded a Twitter Data Grant at the beginning of 2014 as applied to the Jakarta, Indonesian context. This study is a component of the larger research project that can be found at PETAJakarta.org, which primarily focuses on flood disaster management in Jakarta through the use of Twitter.

1.2.1 Jakarta Flooding

Flood management becomes a significant task in every flooding incident due to the destruction that floods can cause in the context of global climate change. The megacity of Jakarta, Indonesia, with a metropolitan population of more than 20 million and rising, is situated in the western side of Java Island, on the north coast of the island. It faces the jaws of the Ciliwung River, the largest river, which divides the city into its western and eastern principalities.

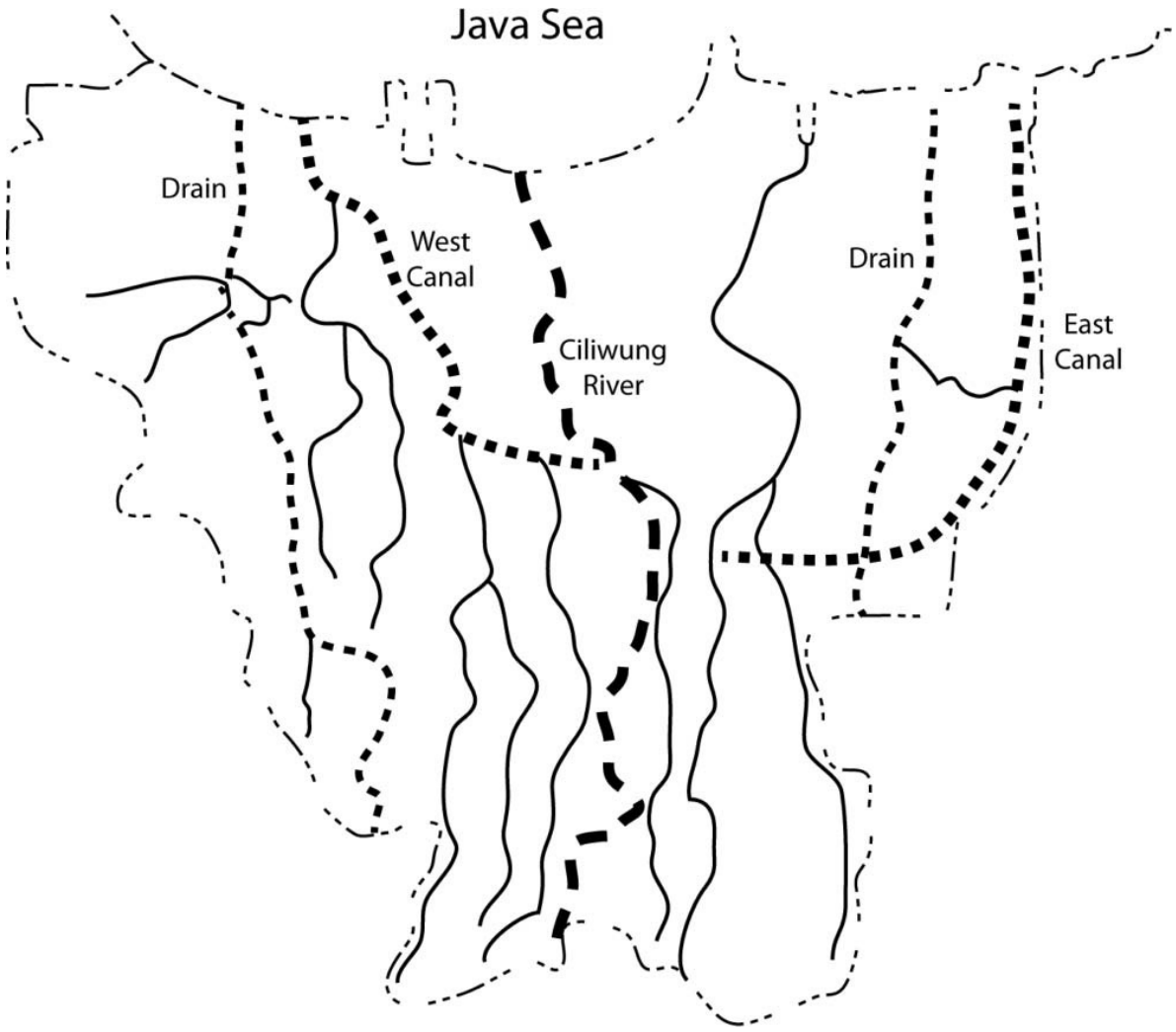

Figure 1. River and drainage ways through Jakarta, Indonesia.[8]

As is demonstrated by Figure 1, there are 13 rivers flowing through the city, but most of them flow northwards toward the Java Sea [8]. Jakarta's climate is tropical and rainy, and is the wettest during the monsoon season, experienced October through March annually. Flooding has always been a problem in the rainy season, due to Jakarta's low topography, climate, congested areas and infrastructure [9]. According to BPBD DKI Jakarta, the emergency management agency of Jakarta, it was estimated that there were more than one million people in the city that suffered from the Great Flood in January 2013. Worse yet, Jakarta residents suffered cascading destructive influences when the civic infrastructure established to cope with the monsoon season partially broke down.

If a one-fold river gate fails, for instance, one of the city's main streets can be destroyed, as well as civic infrastructure sent into chaos. For example, the main transportation hub was closed and blocked by the flood of 2013. Therefore, the bio-system of the whole city begins to feel the pangs of flooding right from the very first rainfall.

In 2005, UNESCO proposed a larger project for Indonesia in an effort to establish a warning system for its archipelago at an estimated cost of $60 million [10]. The system will act to warn authorities of the flow direction of the water, and levels of flooding by location, making it conducive to flood disaster management. Hence, the government is able to simultaneously control the river gates and pumps using the system so they can visualise and direct the flow of the flood to prevent the highest density population areas from being affected and to best protect important civic infrastructure.

1.2.2 The Role of Social Media

The mobile Internet has changed the way people communicate[11]. Additionally, the introduction of social media has meant that individuals can have global reach with their sentiments, opinions, likes and dislikes. The penetration rates of the mobile Internet have increased substantially over the last 5 years. Social media is interactive and a shared information-based network which makes every individual a potential information source. Micro-blogging is the act of sharing short and timely messages, on a broadcast social media platform. Micro-blogging messages may optionally contain geographic information (e.g. longitude and latitude coordinates) leading to the integration of geographic information systems and social media platforms. The integration of social media and geographic information systems is an emerging concept in the geosocial intelligence space.

Micro-blogging now is not only a way to share social messages, but also to grant power to access the data that can be useful during emergency management situations, such as flooding disasters and earthquakes, because the locational information is of vital importance [5]. Locational information can be embedded within a microblogging message using text (e.g. "Jakarta" string) or it can come embedded using the GPS (Global Positioning Systems) sensor on the device being used to microblog.

In effect, any individual who shares a message with a location on it by using a GPS-enabled smart phone *is* a sensor themselves during a disaster [12], such as a flood, which can potentially aid in rescue efforts and providing aid to the target area. When the number of "field" sensors becomes ubiquitous, the distribution and scale of the disaster area can be mapped out and even more specific details shown that would otherwise go undetected. A fundamental example of how GIS can show near real-time information is the Google Maps application that not only provides driving direction solutions and positions in the map but real-time traffic information, by which users can change routes when there is congestion. The traffic information is calculated using Google's servers and traffic algorithms, but real-time traffic data is provided by innumerable users via GPS-enabled smart phones.

Demonstrated uses of social media making use of geographic sensor data was exemplified during the Arab Spring and the Occupy Movement. These events demonstrated the importance of how infrastructure could be used to mobilise citizen resources. Applying this kind of systems thinking to disaster management, given the immediacy of a rising flood, can potentially aid in minimising the loss of life and loss of infrastructure [5]. Hurricane Katrina in 2005 more than any other event triggered the use of new techniques for emergency response. A number of similar case studies have been explored in the emergency management realm and the value of location based services in this realm has been examined [13]. Other notable disaster management events, which made use of social media for the first time in novel ways, were the Queensland Floods[14]. For a given context (i.e. disaster event), all the information is first established by recreating a network visually, in particular a public social network from which an event map is presented.

Geographic information coming from citizens during a disaster event is extremely effective in providing real-time reporting [5]. This kind of crowdsourcing data being delivered by the people for the people is a new data stream known as "volunteer geographic information" (VGI), described above. Geo-social intelligence is an approach used to manage civic problems such as infrastructure management by using VGI from social media. The current geographic information data-source that is built by governments or formal organizations has its limitations. Hence, it becomes a supplement to existing data-sources that are human-powered sensor networks driven by people via social media.

Twitter is one of the most popular micro-blogging social media platforms. Increasingly it is being used as a powerful crowd-sourced tool supporting research in a variety of different

fields. It is forming one of the largest global networks of intelligent mobile sensors. Even in many slum communities, Twitter is already used to capture geographic information about infrastructure and populations.

1.2.3 PETAJakarta.org

It is well known that Jakarta has the highest density of Twitter users in the world. Jakarta is also the location where the government currently has limited or no information about the performance of its hydraulic and hydrological infrastructure systems in the city during flood events. With the pervasiveness of mobile social media in Jakarta, Twitter data can be applied to offer unparalleled insight into the response of citizens and the city's infrastructure to extreme flooding events.

PetaJakarta.org (Map Jakarta), by leveraging the timely data from Twitter's network, is a crowd-sourcing urban data collection project coordinated by the SMART Infrastructure Facility at University of Wollongong, the Jakarta Emergency Management Agency (BPBD DKI Jakarta), and Twitter Inc. The project's primary aim is to assist and improve the lives of residents in Jakarta. It is a web-based platform on which flooding information for Jakarta residents can be gathered, sorted and displayed in real time by utilizing the power of social media. Twitter is providing a framework where all its data can be collected and disseminated by community members via GPS-enable smart mobile devices.

As Figure 2 shows, the entire project PETAJakarta.org contains three parts: CogniCity, Web-based Outputs and Offline analysis. CogniCity is open source software that uses a GeoSocial Intelligence Framework developed by the SMART Infrastructure Facility, University of Wollongong, which allows situational information to be collected and disseminated by community members through their location-enabled mobile devices. It also optimises infrastructure surveys and asset management for governmental actors. Equipped with scalable mapping technology for mobile devices and a critical alert service, this software enables the communication of two-way time-critical information to and from individuals and government agencies. The platform of PETAJakarta.org runs on CogniCity. The web-based outputs use an information display online platform, which presents the crowdsourced citizen data publicly for anyone to access. This research is a part of the "Offline Analysis" system, which is retrospective and not time sensitive.

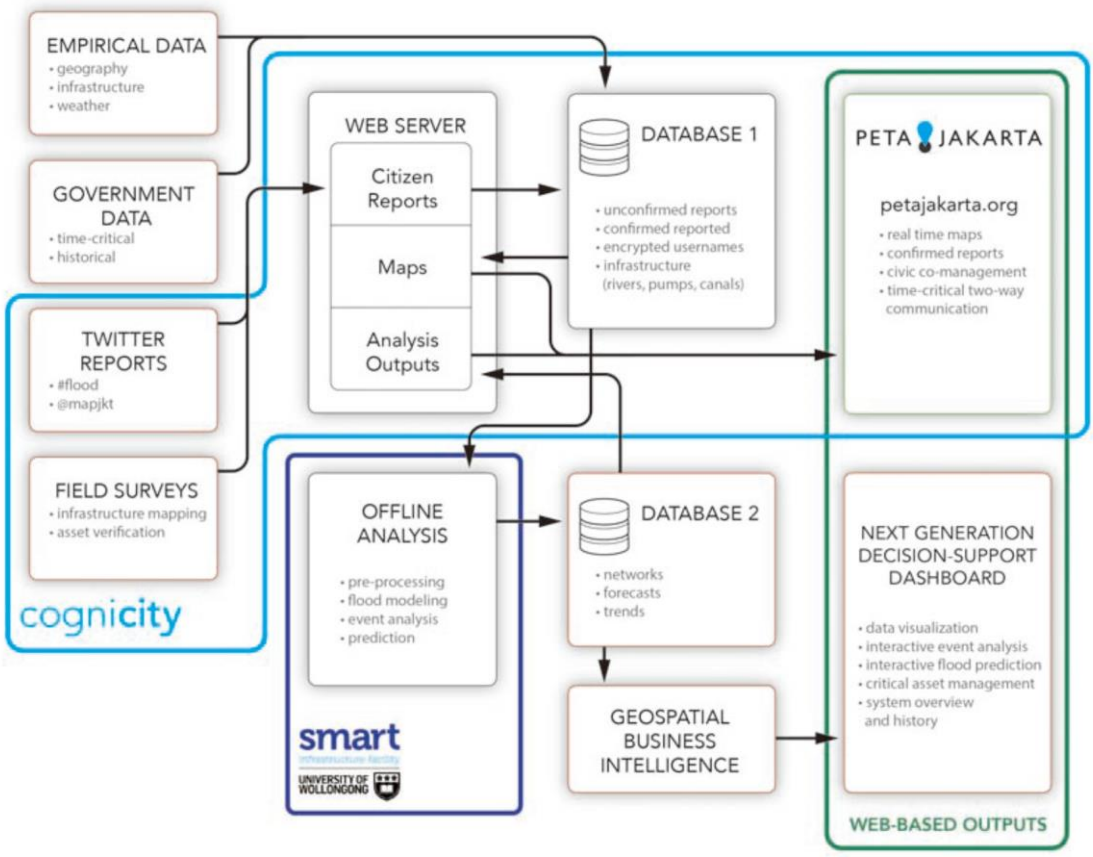

Figure 2. System Diagram for PETA Jakarta.org

1.3 Research Question and Aim

Research question:

To understand the full potential of Twitter as a real-time indicator for flooding, this research project in particular to quantify the temporal relationship between tweets related to flooding, and flood events in the city of Jakarta, Indonesia. Principally the project uses a statistical approach to investigate a time series of tweets related to flooding over specific monsoon periods, to determine whether the number of tweets is proportionally relational to the extent of flooding in the city at that point in time. Therefore, in this study, the focus is on the demonstration of the relationship between Twitter activity and flood events using statistical means. Past Twitter data and the real observed flooding events are used as the basis for modelling the urban flooding event in its totality.

Research aim:

The overall aim of the PETAJakarta.org project is to enhance the capacity to understand and promote the resilience of cities to both extreme weather events due to climate change and to long-term infrastructure transformation with the process of climate adaptation. It is hoped that such research can contribute to a wider understanding of the potential of social media to act as a real time crowd-sourcing tool during extreme weather events. The long-term goal of this research is to quantify the utility of social media data during flood events within the context of a civic co-management framework.

1.4 Research Objectives

This research is divided into a series of data experiments that will use statistical tests to examine different aspects of the corpus of Tweet data that was collected.

To advance the comprehensive understanding of the relationship between flood events and Twitter activity in Jakarta, this study will quantify the relationship by addressing a spatiotemporal comparison of historical tweets with official flood data from the 2012-2013 and 2013-2014 monsoon seasons. The study relies on three stages of enquiry, each of which corresponds to a specific experiment.

As such, there are distinct objectives for each of the three stages:

- **Stage 1-Summary Statistics:** To provide an overview of the data
- **Stage 2-Time Series Plot:** To explore the relationship between Twitter activity and flood events
- **Stage 3-Relationship between Twitter activity and Flood events:** To test whether there is a relationship between number of tweets and number of flooded areas over time

The first two stages can be regarded as the pre-processing stages in the processes of analysis and modelling, which are attempting to summate all data and gain a preliminary understanding of flooding events and Twitter activities beyond the data itself. Stage three seeks to provide a comprehensive analysis of data and relationships between flooding events and Twitter activities.

1.5 Outline of the Thesis

This thesis employs a traditional layout to investigate urban flood modelling using geo-social intelligence. It begins with defining the problem through a thorough literature review of previous work in the interdisciplinary domain of GIS and social media from 2009 until September 2013. Specifically, the literature reviewed concentrates on the adoption of Twitter in broad event detection and emergency management contexts using crowd-sourced civic data to grant organisations the ability to develop preparedness and response strategies. This thesis is an integral component of the larger Cognicity architecture which is now available through the Open Data Institute (ODI). Chapter 3 uses the analysis of previous literature to address the gaps found in methodology, namely the ability to interrogate big data using statistics- from descriptive to correlational analysis. A detailed description of how the experiments were set-up and the software used to manipulate the datasets is included in this chapter as well as the main hypotheses. The expected outputs include tabular and graphical results working at the Twitter-based "Tweet" unit of analysis. Chapter 4, includes a thorough description of the results of the three experiments completed in this project, providing further insights into the relationship between tweets in the urban megacity of Jakarta and urban flooding hotspots during seasonal monsoonal activities. Chapter 5 concludes with the major outcomes and future work.

Chapter 2. Literature Review

2.1 Introduction

2.1.1 Technical Background

The problems of the distribution and acquisition of crowd sourced information have been addressed by emerging technologies. As more and more mobile devices are equipped with GPS sensors, computers and Internet connection with advanced server- and client-side technologies, users can actively participate and be satisfied with these applications and location services. From a distinct perspective, the user is becoming a complex stakeholder given the dual role held as producer and consumer. Prosumers, as they are now known, have an important role to play in society. Web 2.0 has enabled and encouraged citizen-generated reporting useful to problems requiring large-scale coordination. Participation by citizens in once government-only problem solving is a completely new paradigm that has been enabled by emerging technologies. The act is known as "participatory sensing".

Citizens wishing to engage in contributing vital flood knowledge using Twitter do not require previous expertise. They simply go about their business as usual and may additionally opt to include particular hashtags that are encouraged for ease of near real-time data mining. Goodchild, in 2007, named this practice as "Citizens as Sensors", where Volunteered Geographic Information is created, gathered, and spread by those individuals or groups who can use Web 2.0 [4].

The interactive networked and shared model of "People as Sensors" information is supplied for free and entirely voluntary. Haklay calls this new social web mapping application the evolution of the Geo-Web [15].

Social Networks are an important part of this development, combining new information with communication tools and applications, attracting hundreds of millions of users. Boyd and Ellison point to the term Social Network Sites (SNS), on behalf of individuals who construct an online profile communicating with other users, in order to share their common ideas, activities, events interests and backgrounds [16]. Furthermore, Location Based Social Networks improve existing social networks, adding space with location services (for more information, see [7] and [6]). For example, users upload geo-tagged photographs from Flickr,

checking in by a venue with Foursquare or commenting on a local event on Twitter. These are all digital touchpoints that leave behind digital traces. Geo-information drawn from the Location Based Social Network is included under the umbrella of volunteered geographic information, although sometimes users themselves do not realise they are leaving behind these breadcrumbs.

However, Harvey argues that a preferable term would be "contributed" data, since people do not consciously volunteer their data, but use the platforms to generate it for their particular purpose [17]. When data generated for one purpose is used for another purpose, no matter how honourable the aim, there are privacy and ethical implications that come to the fore. While outside the scope of this project, the utilitarian approach has been espoused here- for the sake of the common good this data can help aid Jakarta's securitization.

As for Twitter, users can publish short status messages with at most 140 characters and may attach photos and videos. The act of sending a message using Twitter is known as "tweeting". These status updates may contain syntax such as hashtags, which can refer to a key word or jargon relevant to the given topic the users are discussing or commenting about. Users have the option of "following" other users, or being "followed" themselves. One can tweet or retweet or even favourite someone else's tweets. Additionally, one user can send direct messages to another, and any user can search the entire corpus of tweets for specific information.

According to Twitter, an average of 500 million tweets is being generated per day among about 270 million monthly active users. With the permission of the user, each tweet requires geo-location information from the GPS sensor within users' devices. These location structures allow users to exchange details of their location as an important interaction via the Internet. Location Based Social Networks connect our physical world and network services containing three layers of information: user layer, location layer and content layer [18].

Therefore, a status update which users publish using Twitter represents a spacing signal in a semantic information layer. After registration in Twitter, all tweets can be recorded in real-time through the streaming API. The Twitter API prompts for allowance filtering of selected tweets or a choice to access only those tweets obtained by geo-referenced Twitter messages in a bounding box. From this spatiotemporal information layer, we can see that a byproduct of individual social interaction may drive research in spacing structures. In the last five years,

there has been an increase in papers addressing location based social networking. The following literature review looks to identify previous works that are relevant to this project in application, aim and method.

It will be particularly interesting to view the emergence of the use of Twitter for research purposes, and specifically disaster management, in this case flooding. It is important to note, given the nature of this project, that the review will draw on interdisciplinary work.

2.1.2 Existing Literature Reviews

Given the nature of this research, this literature review was focused on searching in the following journals: *International Journal of Geographic Information Science*, *International Journal of Remote Sensing*, *Photogrammetric Engineering Remote Sensing*, *Computers and Geosciences*, *Transactions in GIS and Geo-Informatics*. Only the top-ranking GIS journals were selected for the literature review. Interestingly, besides literature surveys and basic non-systematic reviews, no journal articles were found conducting a systematic literature review.

This primary observation emphasises the need of further research that focusses on systematically reviewing literature in the geographic information science (GIScience) domain. Related to this field of study, Horita, et al. estimated the actuality of search for a conference paper that analyses VGI for disaster management and applied a systematic literature review, including a screening process of important electronic databases [19]. Roick & Heuser offered a general review of current research on Location Based Social Networks which was not systematically conducted. However, they pointed out the need of further studies that investigate how social networks can be applied to special use cases [20]. Blaschke & Eisank also conducted a non-systematic keyword based literature search, which compared the “GIS” and “GIScience” and their total number of citations over time [21]. However, existing literature reviews of GIScience have been performed in a non-systematic manner, lacking any descriptive statistics of the field at large.

2.2 Review method

This review will follow the guidelines given by Kitchenham & Keele in 2007 for conducting a literature review, dividing the research into three main sections[22]:

1. Planning the review,
2. Conducting the review by electronic databases, and
3. Reporting on the final results.

The flowchart review model in Figure 3 presents the adopted workflow approach. The following paragraphs and sections are divided according to the review process shown in Figure 3.

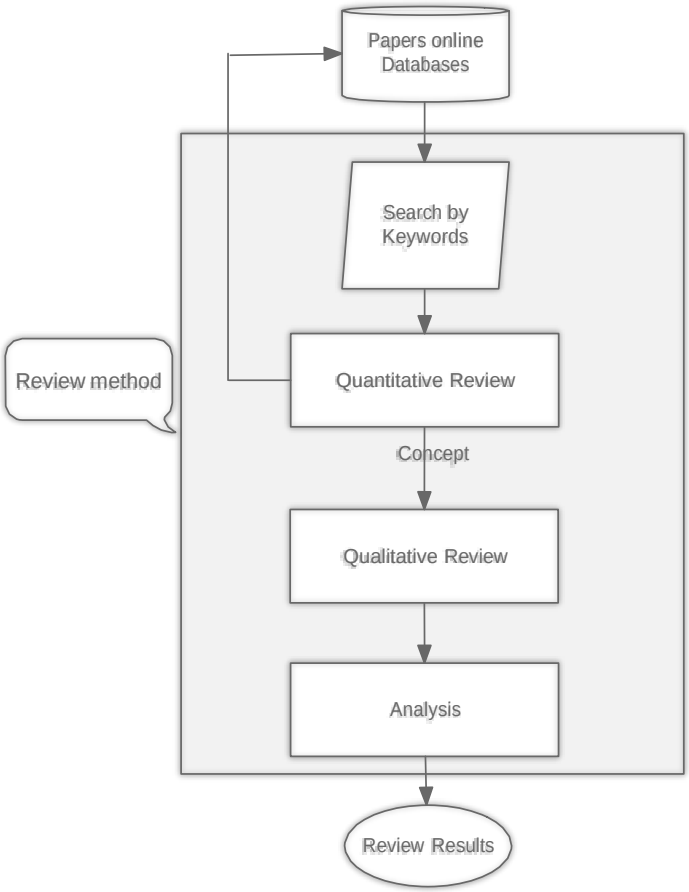

Figure 3. Literature Review Workflow Process

The detailed review method steps are shown in the shaded box in Figure 3. For identifying primary studies presenting detailed state of the art research, a clear research question was established, followed by the selection of eligible literature sources by:

- Choosing journal, workshop and conference proceedings published between 2005 and September 2013 in English (this is consistent with the time in which the study was conducted).
- Choosing multiple digital libraries related to information identified by Brereton et al. in 2007 and more with GIScience relevant digital libraries [23].

The databases used for searching relied on defining a set of keywords and searching all published papers before thirty-first of September 2014. To ensure that there was limited bias, a test review with preliminary trial searches was carried out using defined search strings for the data extraction process. Initially, 288 papers were found, while 92 were finally used in the literature review. It should be noted that duplicate "hits" were found between various electronic databases. Ultimately, during the paper screening process, 42 papers were chosen in order to demonstrate the linkages between the formulated research questions. 15 of these final 42 chosen papers did not address their methodological approach in analysing Twitter-based data.

2.3 Review Results

The number of papers related to Twitter specific examples increased substantially in the review period. Between 2009 and 2012 the quantity of published papers that directly referenced Twitter as a microblogging engine as applied to organizational preparedness increased from 27 to 84.

2.3.1 Methods Used to Analyse Data from Twitter

A deeper examination of how Twitter was utilized within the selected review papers indicates how the social media application was utilized by organisations. The methods demonstrate that Twitter data was utilized as a data input. About a third of papers utilized all the information layers including the Tweet message, the geo-tag and the timestamp. These papers were mainly about spatiotemporal and semantic analyses. About a tenth of papers focused on researching spatiotemporal in Twitter but did not include any form of semantic analysis.

More than half of the paper did not address spatial information requirements but only considered the semantic information of the tweet itself. They analyzed the content of tweets and constructed a semantic network to get more non-spatial posts with geographic information to deduce people's location. Among these papers, 4 papers only analyzed the Twitter posts to get location and mark landmarks from textual information. Besides, one paper [24] deeply analyzed semantic tweet frequencies to distribute and locate non-geo-tagged tweets to those with a geographical reference. About 10 papers added some analysis of follower and following activities of Twitter users, and 5 papers added a label and 2 papers included a URL analysis. Metadata that describes users and their personal information was found to be a key outcome of conducting a Tweet metadata analysis. This user-centric focus was applied within 6 of the reviewed papers, includes the analysis of Twitter profiles metadata and tweet posts as well as social relationships, for the prediction of user locations and the clustering of similar users.

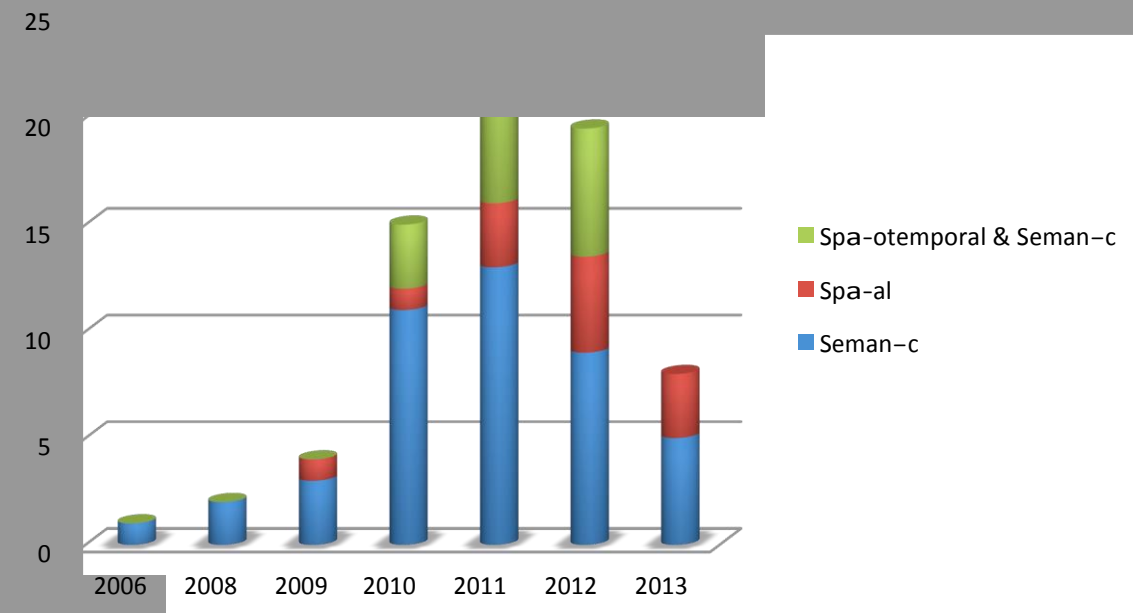

Figure 4. Paper counts per year

Most reviewed papers between 2006 and 2011 conducted research on Twitter by using non-spatial information to study the temporal evolution of information in Twitter, as shown in Figure 4. At the same time, only one reviewed paper in 2009 focused on researching Twitter data with spatial information. Therefore, the number of reviewed papers that made use of spatial information peaked in 2012 with a steady increase in proceeding years. As the number

of papers focusing on spatial aspects of Twitter data increase, the number of reviewed papers researching spatiotemporal and information grew also. There are a number of reasons for this, including the fact that many smartphone apps now have the location feature enabled by default.

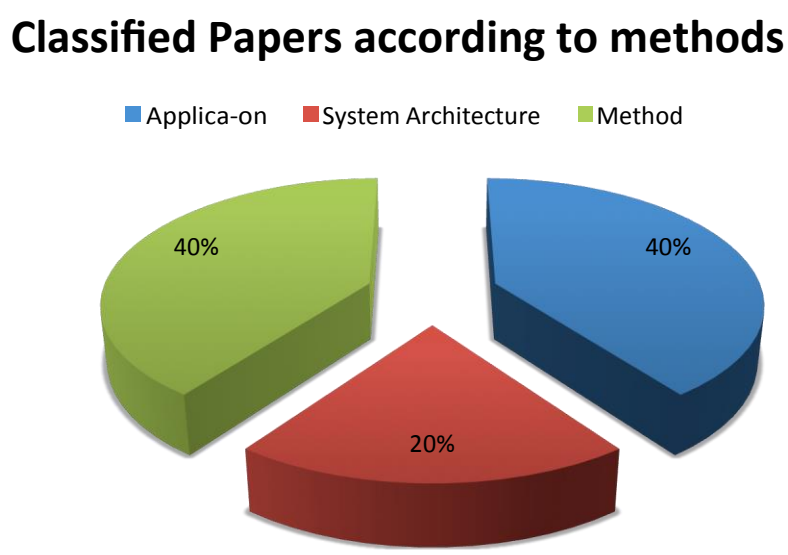

Figure 5. Classified Papers according to applied methods

We can see from Figure 5 that 40% of the articles reviewed had a significant technical component which focused on investigating and developing methods of exploring, extracting, validating and aggregating Twitter data. One fifth of the review articles went further, providing a conceptual model which could be implemented using a system architecture so that researchers could collect and process data from the Twitter streaming API. The rest of the review papers were mainly about the application aspects of Twitter. Looking deep into the applied methods, more than 92% of the 55 papers investigated methods for event detection in Twitter. Those methods and approaches analyzing the social network of Twitter to get location are also frequent methodological applications. Four of the review papers were working on topic detection and about 11 papers had no specific identifiable method.

2.3.2 Management of Health

Twitter, together with its contents, was identified as a method to estimate market prices for medicine when a flu pandemic happened [24]. Sofean & Smith, in 2012, hold the view that disease reports from this social media tool formulate an ontology of medical terms classified by SVM standards [25]. In addition, keywords from tweets were consolidated and utilised to handle semantic similarities and spatiotemporal events during the dengue fever in Brazil [26]. What is more, Lampos & Cristianini in 2010 conducted a mathematical study in UK to seek several inner relationships between Twitter posts and real world disease reports through a correlation regression assumption [27].

2.3.3 Monitoring of Traffic

Without a doubt, tweet information could be adopted to manage the operation of the traffic systems. In order to match traffic conditions from Twitter about road networks in Thailand, online content was found to be valuable in determining the spatiotemporal traffic-related information with the support of NLP and keyword filtering method [27]. Through the usage of gazetteer, Gerais et al. in 2012 conducted research, detection and locating for traffic incidents with Twitter [28]. Sakaki & Matsuo in 2012 also operated an investigation in Japan with additional classifications of driving data from the social media to achieve whether close connection exists [29]. Similarly, with the help of NLP, traffic-related information could be gathered from Twitter and finally be applied to judge the plausibility of events [30]. It is clear that research in this field is focussed on mining typical or characteristic motion patterns from a single user or collection of Twitter users.

Wakamiya & Lee, in 2012, extracted Japanese driving data from tweets by spatial partition methods such as administrative parts or a grid cluster [31]. Similarly, Ferrari et al. in 2011, together with Fuchs et al in 2013, extracted urban patterns in the USA by spatiotemporally investigating Twitter and its user activities [32]. Yuan et al. also developed a program to analyse location and user activity on Twitter for forecasting mobility patterns. Sadilek et al. mined data of spatiotemporal behaviour curves regarding Twitter users [33]. Furthermore, in 2013, Andrienko used Twitter data to assemble, classify and analyse spatial distribution, in order to determine spatial behaviours [34].

2.3.4 Making Inferences Using Location

Location inference refers to the procedure of retrieving and mining location-related information from Twitter, with not only metadata, but also the simple tweet contents. To attain sufficient geo-location and geo-referenced tweets, Gerais et al. in 2012 deduced locations through user profiles and their net friends, while Finin et al. constructed a named-entity recognition program on Twitter by establishing a sourced natural language processing, based on a language-based mode identified by Kinsella et al. in 2011 [28, 35]. Hecht et al. in 2011 estimated the semantic geo-referencing method based on user content derived from Twitter, by comparing term frequencies with Naive Bayesian Classifier [36].

Kulshrestha & Gummadi determined user geo-location through analysing user origin and Twitter population. Li et al. tried several other approaches such as a ranking method for concluding POI tags on tweets in 2012. GPS and related devices are also adopted to indirectly evaluate the geo-location from Twitter contents [37]. In addition, Gonzalez & Chen, together with Hiruta, et al. in 2012, utilised profile location and semantic classified tweets to create a site inference system [38]. Concentrating on a Twitter data analysis, Watanabe et al. in 2011 developed rules so that computers could automatically tag non geo-referenced Twitter data for local events with actual location [24]. Furthermore, Dalvi et al. in 2012 identified the geographic information of Twitter users through connecting indirect spatial data to real world spatial objects [38].

2.3.5 Social Network Analysis

Social network analysis focuses on features of users within a network and their operational behaviours. Recent scholarship concentrates on content information from Twitter posts. Based on the linguistic test on Twitter, led by Hong et al., researchers identified that 51% of Twitter tweets are written in English. In addition, after the adoption of the LDA model, based on several linguistic characters, experts concluded how political and ethnicity issues, together with some affinity aspects, influence the unique activity of each Twitter user [39]. Wu et al (2011) also point out that the affinity of users for diversified events can be differentiated due to their characteristic lifespan.

Similarly, Takhteyev et al. applied several algebra techniques and flight information to investigate the relationship between particular spoken languages and the social ties of Twitter users [40]. Through research on user tweet and re-tweet behaviours, experts have conducted studies on the perceptions of individuals with respect to particular social topics [41]. Weng et al. conducted a study on the impact of diversified users on calculating and ranking topic similarities, based on LDA and the relationship structure [42]. With the research conducted by Krishnamurthy & Arlitt in 2006 and Yardi & Boyd in 2010 [43], it has been concluded that distinct groups of Twitter users behave in a specific manner based on their usual social network conversations. What is more, Cranshaw et al. evaluated Foursquare data from Twitter, with a spectral clustering algorithm, and found that characteristic neighbourhoods exhibit their own spatial and social proximities.

Sentiment and emotion investigation on Twitter is another breakthrough point whereby researchers conducted an analysis of social network and computational linguistics. For instance, Maximum Entropy machine learning technologies [44]. In addition, Wang et al. adopted a system to conduct real-time Twitter Sentiment Research during the US election. Therefore, mining emotional vocabularies could be used as an alternative method to determine the relationship between sentiments, topics and social ties [45].

2.4 Discussion

2.4.1 Synthesis of Relevant Literature

Before conducting the literature review, it was evident that the growth of publications concerning Twitter between 2005 and 2013 was significant. In fact, it is obvious that attention from users to Twitter has increased, so too has the attention that Twitter has received from researchers. There is a multiplicity of reasons for this, including the movement toward crowdsourcing government and non-government open data, and harnessing big data strategies toward environmental sustainability, citizen science initiatives, and advancing humanity for good strategies. During 2005 and 2010, ACM was the electronic database where most studies reviewed were published. However, after 2010, there were publications about Twitter from a variety of publication outlets. As a result, related investigations have widely spread due to the difference of target audiences of each electronic database. Since the review was conducted,

the number of publications has grown exponentially, indicating the importance of social media toward solving big societal problems. Increasingly, the lower cost of sensors within an Internet of Things infrastructure, also has had a major impact on the way that Twitter might well be used in the future to generate data, machine-to-machine.

A large number of studies were focused on spatiotemporal Twitter analysis (43%) that described the collection and handling of textual data sources from tweets through keyword filtering. It is clear that limitations exist because those studies were focussed on the uncertainty and sparseness of the data, which made it difficult in validating and comparing reference data. What is more, the Twitter API query also negatively influenced the number of tweet posts.

Location Based Social Networks and Twitter derived from the field of computer and information sciences, and interestingly, the dominant topic of academic journals was Twitter from 2005 to 2011. However, these studies demanded new categorisation to be incorporated into the review of literature beyond the central theme of the thesis. This involved the integration of dimensions or disciplines such as earth-geoscience and social science. Currently, consumers are more likely to exchange location information on a mobile device or platform equipped with GPS capabilities, which could underpin the penetration effects and usage of social network. This in effect results in the development of the geoscience field, whereby the goal is to utilise “Citizens as Sensors” for forecasting when and where a natural catastrophe happens [4]. As a result, adopting spatiotemporal information from location based social networks, like Twitter, creates many research opportunities in the field of GIScience.

Table 1 provides a summary of the results of reviewed studies, presented in chronological order. Geo-referenced data from Twitter offers accurate location information that can be used in many ways. High spatiotemporal reliability is valuable in many disaster and emergency management situations, such for real-time detection and analysis of earthquakes. In addition, the information can also be used for disease and health management.

For instance, these data can indicate the spatiotemporal contagion situation of illnesses, compared with the official data. Given the up-to-date, cheap, often free and potentially widespread nature of the information, companies are also able to develop spatiotemporal solutions for their particular benefits. An example would be early-warning systems,

delivering spatial and real-time information regarding issues that may affect the business operation. The data is also beneficial to monitoring of traffic and human mobility, whereby the can be adopted to achieve rapid reactions and better managerial outcomes.

Date	Authors	Research Method	Research Overview
2009	Longueville & Smith	Geographic-feature-based extraction from tweets with keywords to land mark	Spatiotemporal Tweets accurate to real world event include indirect geographical information and URLs tend to media
2010	Lee & Sumiya	Central Points of K-means cluster used to form voronoi diagrams	Unusual crowd activities assuming abnormal events, such as earthquake, have been detected by observing geographic regularities within defined regions. (Case study on Japan)
2010	Lamps & Cristinini	Matching geo-referenced tweets within 10km radius, n-gram textual analysis	Significant correlation at 95% between the flu epidemic related tweets with the official health report. (Case study on UK)
2010	Sakaki et al.	Using SVM to classify tweet locations into text by Kalman filtering	Estimation of the earthquake location and typhon track by tweets is possible.
2011	Maceachren et al.	Filter aggregated grid based number of geo-referenced tweets with a set keywords	Approach was able to extract and validate locations of tweets during an earthquake event
2011	Earle et al.	To detect spatial outliers by tweet frequency analysis filtered with spatiotemporal	It is to compare worldwide with Twitter data with official geological surveys on earthquake detections. There are only 48 earthquakes have been

Date	Authors	Research Method	Research Overview
		keyword	detected within Twitter in 5175 earthquakes with the average 2 minutes detection delay.
2011	Stefanidis et al.	Spatial hotspot detection	Geopolitical events (e.g. riots) and hotspots of other crises have been detected and information dissemination within Twitter studied to improve the situation awareness and emergency response.
2011	Veloso & Ferraz	Filter tweets with a set of keywords	Strong correlation (at 95%) between spatiotemporal distribution of tweets related to dengue fever cases and official statistics(case study in Brazil).
2011	Wanichayapong et al.	Geocoding of geo-referenced tweets to road related attributes, Tweets have been filter with a set of keywords	Point and link based traffic incidents from Twitter have been classified into road segments with 93% accuracy and on points with 76% accuracy.
2011	Li et al.	POI matching and ranking method	The developed ranking method predicted the POI tag of tweets bases on textual information and time (case study in Chicago).
2012	Terpstra	Mapping of geo-referenced tweets which have been filtered with a set of keywords	Extract event information for storm and demonstrate the insights for improving disaster management and relief (case study festival in Belgium).
2012	Chae et al.	Seasonal-trend	In 3 case studies, events have

Date	Authors	Research Method	Research Overview
		decomposition of LDA semantic topic modelled tweets to detect abnormal spatiotemporal pattern	been detected by using location information and textual information
2012	Kling et al.	Spectral clustering and geographical heat maps of LDA semantic topic modelled tweets	Temporal patterns and functions of urban areas have been detected
2012	Boettcher & Lee	Keyword frequency analysis of DBSCAN clustered tweets	There is 68% precision of events' been detected by estimating the average tweet frequency of keywords per day in and around a potential even area
2012	Lee & Hwang	Text based grouping method correlating geo-referenced tweet with user set profile location	Correlation of user profile locations and geo-referenced tweets showed that more than half of all tweets are posted in the user's hometown. 30 % of Twitter user did not have any posts near their set profile location.
2012	Hiruta, et al.	Classification of geo-referenced tweets called Place-triggered geo-referenced Tweets. Tweets have been filtered with a	Tweets have been successfully classified into type of places (whereabouts of people, food, weather, back at home, and earthquake). Detection of place triggered geo-referenced tweets

Date	Authors	Research Method	Research Overview
		set of keywords	had 82% accuracy.
2012	Dalvi et al.	Probabilistic Distance based model with parameter inference using EM algorithm. Tweets have been filtered with a set of keywords	Language and distance based model was able to infer and match tweets with a real objects geographic location (example POI restaurants)
2012	Cranshaw et al.	Spectral clustering of geo-referenced check-ins posted through Twitter. Activity have been classified according to check-in venue categories	Social media check-ins and qualitative interviews revealed collective social behavior of people differentiating a city into "Livehoods" which correspond to municipal boundaries (case study in Pittsburg of USA)

Table 1. Study overview of papers on spatiotemporal Twitter analyses

Nowadays, it is obvious that experts can extract, mine and even forecast users' locations information by using Twitter data. As a result, academics could use this data to conduct detection and other managerial tasks through information from Twitter and other metadata. However, Twitter data is mostly used in the United States. For example, although Brazil has one of the highest use-density of Twitter, related data are only adopted for two fields. Nowadays, most studies do not pay attention to the quantitative geographical spread of Twitter information, so that it may make it difficult in achieving other outcomes in other research fields.

One of the major uses of tweet information is in disaster management, and most studies are focused on the field of earth-geosciences to deal with emergent disaster monitoring. Since several studies have started to focus on event detection, new applications have benefited from

the related theoretical contributions and analysis, especially in fields with limited availabilities of sufficient official data.

2.4.2 Gaps & Limitations in Past Research

Concentrating on current related research, one significant research gap in the GIScience domain is the limited utilisation of common methods, such as spatial data mining technology, when experts want to adjust to recent data sorts such as uncertain and sparse geo-referenced social media feeds. Although the use of density based spatial clustering techniques and point based observations, two clustering methods benefit studies. The uncertainty of data, dynamic and diversified point densities and geographic scale effects, still do harm to the measurement of parameter value for distance. As a result, the procedures in the studies reviewed could not completely be incorporated into the real world of geographical-based datasets [46]. However, great potential still exists in mining and processing data from Twitter information for improvement.

A little less than half of the papers were focused on event detection, and only one fifth of the reviewed investigation formulated several system architectures or some potential service applications. The PetaJakarta project differs substantially from these studies because it allows data to flow through Cognicity with an end point for data visualisation and immediate application toward emergency preparedness and response. This fresh analysis method, which can spatiotemporally mine data from social networks like Twitter in near real-time, is destined to open a new door for researchers in a variety of disciplines.

It is clear that keyword-filtering in microblogs are widely used as more than three quarters of the research reviewed extracted spatiotemporal information from Twitter in this way. In addition, the increasing adoption of computer linguistic approaches to deduce textual messages from Twitter, together with associated spatiotemporal analysis, is presently an emerging field which is currently still lacking real-world implementation beyond small scoped/scaled studies. It is one thing to extract 500 hashtags in a small geographic location about a flood, and another to get more than a hundred thousand tweets for a monsoon from a dense urban area (e.g. #banjir) in the span of a weekend. Due to the wide application of semantic data from tweets, one is able to link theoretical studies to practical analysis of semantic and spatiotemporal information. It is true that the primary adoption of those textual

data, such as user profile, follower and following information, is for investigation of their social networking and research about user emotions, but related data layers must not be ignored. Experts who focus on those investigations could benefit from the utilization of spatial data sources of geo-tagged Twitter messages. What is more, several researches also conduct programs to trace how information of events spreads and transforms social networks by investigating associated website links. There is no doubt that the technology and theories could be useful to manage lots of events or incidents, like illness, catastrophe and business development.

In conclusion, GIScience and its effects seemed to have been limited until September 2013, especially in the analysis of spatial methods. One of the major issues is that only 7% of the papers reviewed were conducted by researchers with a geoscience background but in the last few years especially, there has been a significant shift in researcher skillset to see the integration of GIScience and IT/social media expertise. Although the application of Twitter information to determine location is studied by many researchers, the lack of GIScience qualifications has created a research gap. In summary, four main limitations of current research are identified in this GIScience study:

- Firstly, there are a limited number of traditional approaches or techniques to spatially handle and deal with information from location-based social networks due to its uncertainty.
- Secondly, experts currently could only marginally investigate the geographic scale effects of information source from tweets.
- Thirdly, most reviewed research on the spatiotemporal analysis of information from Twitter only uses one or two methods. As a result, limited combined effects were shown or found from recent studies.
- Forth, limited research has been conducted to find methods to upgrade data mining filtering perspectives for information from Twitter.

Besides the angle of GIScience, current researches also own limitations due to their use of several terms that could be unclear or have heterogeneous understanding in diversified academic areas. In addition, some researchers arbitrarily decided the keywords of their studies of spatiotemporal analysis of Twitter data so that several different vocabularies are used as keywords, which added to the complexities of the literature review [47].

2.5 Conclusion

With the support from current academic sources, a systematic literature review was conducted in reviewing present scholarship, related theories and methodologies pertaining to the use of information from Twitter to conduct spatiotemporal processes, such as event detection. To complete this review, a large amount of diversified online libraries and databases were searched, in order to achieve a large and sufficient amount of relevant academic information. To avoid the negative effects of bias during the research, an iterative keyword searching method was adopted. In addition, with the support of both quantitative and qualitative review approaches, the searching process reduced the proportion of papers
that were not checked. With the systematic literature review, investigations of Twitter, from a new angle, as a location-based social network with statistically-based analysis could be performed. What is more, new doors for the investigation in the field of GIScience were opened by researchers who combined GIS with location-based social networks. Without a doubt, with the methods offered by GIScience, the spatiotemporal extraction and analysis of real-time social-media information, such as data from Twitter, could be much more convenient and effective.

Chapter 3. Methodology

3.1 Introduction

This chapter uses the gap found from the literature review in chapter 2, and addresses the need for conducting an experiment with social media data for civic co-management strategies towards the PetaJakarta Twitter Grant project. A deficiency was found in literature published prior to September 2013, whereby a lack of integration between GIS and social media existed, as well as scant techniques on approached to analysing the linkage between location information in microblogging content and related metadata. This thesis attempts to demonstrate the importance of geosocial intelligence toward urban flood modelling and management, specifically in Jakarta. This chapter sets out the chosen approach to modelling crowdsourced data to prove in the strong relationship between the social media posts by everyday citizens as per identified flood zones. The chapter includes a research strategy, the data sets available for the study and how they were extracted, the systems configuration for the study, the actual experiments conducted in the thesis, and the manner in which data analysis is to be conducted.

3.2 Research Strategy

This research employs the quantitative approach to research, using statistical, mathematical and computational techniques. Quantitative data is any data that is in numerical form such as statistics and percentages. The objective of the quantitative method, in this instance, is to develop and utilise mathematical models, theories and hypotheses pertaining to the phenomena in question: Jakarta Flooding events. Twitter data (from tweets) containing semantic and spatiotemporal information was gathered, filtered and settled, after which the data was analysed statistically. Given that this research is focused on models of urban flooding and flood management, the quantitative technique is ideal in that it provides the fundamental connection between empirical observation and mathematical expression of quantitative relationships.

3.3 Multiple Datasets

Twitter data was obtained for two particular monsoons, and was filtered by relevant keywords in Bahasa Indonesian or English. For instance, keywords such '#banjir', '#flood', 'genangan', 'pool', 'terendam' and 'submerged', were supplemented with observational data of Jakarta flooding events as recorded by the Indonesian Government. The latter contained data such as time, coordinates and count, all of which were recorded in numerical form.

Two datasets were used in this research, as described in the following sections.

3.3.1 Dataset A

The initial database contains two tables: "floods_2012_2013" and "floods_2013_2014". The respective flood tables comprise polygons indicating flooding extents at a given point in time. The corresponding tweets tables contain tweets at a point in time and their location meta-data, if available. Tweets were selected if they included geo-location data and were either within the following bounding box: bounding_box:[106.5894 -6.4354 107.0782 -5.9029] or the user's bio-information contained the word 'Jakarta' (bio_location_contains: "Jakarta"). Thus, the archive contains both spatial and a-spatial tweets.

Real-time tweet data from Twitter #DataGrant and UTC Time Frames:

00:00 30/10/2012 – 00:00 02/3/2013 (Monsoon 2012-2013)

00:00 30/10/2013 – 00:00 02/3/2014 (Monsoon 2013-2014)

3.3.2 Dataset B

The second dataset of Jakarta Flooding events (areas) contained data provided by the Indonesian Government, namely the agency BPBD DKI. This included:

1. archive of flood tweets for the 2012-2013 and 2013-2014 monsoon seasons
2. archive of flood extents for the 2012-2013 and 2013-2014 monsoon seasons

Areas shown are at the 'RW' municipal scale. At this scale, it is probable that the entire area shown is not flooded, but data for the next scale down 'RT' is not available. Only the 'RW' areas, which were flooded for the given time period, are shown.

Generally, the data shows the 'RW' areas which have been marked as affected by flooding over a 12 hour period up to the specified point in time (normally either 06:00 or 18:00), although this frequency may increase and decrease depending on conditions. Therefore, a

record marked as 06:00 shows the areas marked as affected by flooding from 18:00 on the previous day.

3.4 Systems diagram

3.4.1 Framework

The open source software CogniCity was employed in this study in order to analyse the collected data. CogniCity is a Geo-Social Intelligence framework, which runs on the PostreSQL object-relational database management system and uses the PostGIS extension to support geospatial data, including the locations of Tweet reports.
In this research, offline analysis and modeling processes are based on the Cogni-City framework, using the data from the Cogni-City DataBase, with the results feeding back into the Cogni-City workflow. Cogni-City contains three main modules: cognicity-server, cognicity-reports and cognicity-web. The central module of the whole framework is cognicity-server which provides the main database server and HTTP API, and can also serve up static web content.

Interacting with the Gnip Powertrack API to gather reports in real-time, the cognicity-reports module can determine the tweets report's original position by using Twitter's location services. In addition, it can confirm the report by messaging back their users. In the project PETAJakarta.org, most of the functionalities can be replicated by providing a set of HTML templates from the cognicity-web module, as to quickly set up web reporting of Cogni-City data. The offline analysis will be reported in cognicity-web through the cognicity-server module in multiple manners, such as plotting, numbers, and polygons.

3.4.2 Research Diagram

As mentioned earlier, two datasets will be processed within the closed loop research architecture of CogniCity. Dataset A flows from the cognicity-report module for pre-processing in the Cogni-City DataBase. Dataset B, the seasonal observation report of flooding events provided by the Indonesian Government, including the number, the time and the positions of flooding areas, will also flow through the Cogni-City architecture Figure 6 illustrates the Offline Database including Dataset A and B, on which the experiments will be based.

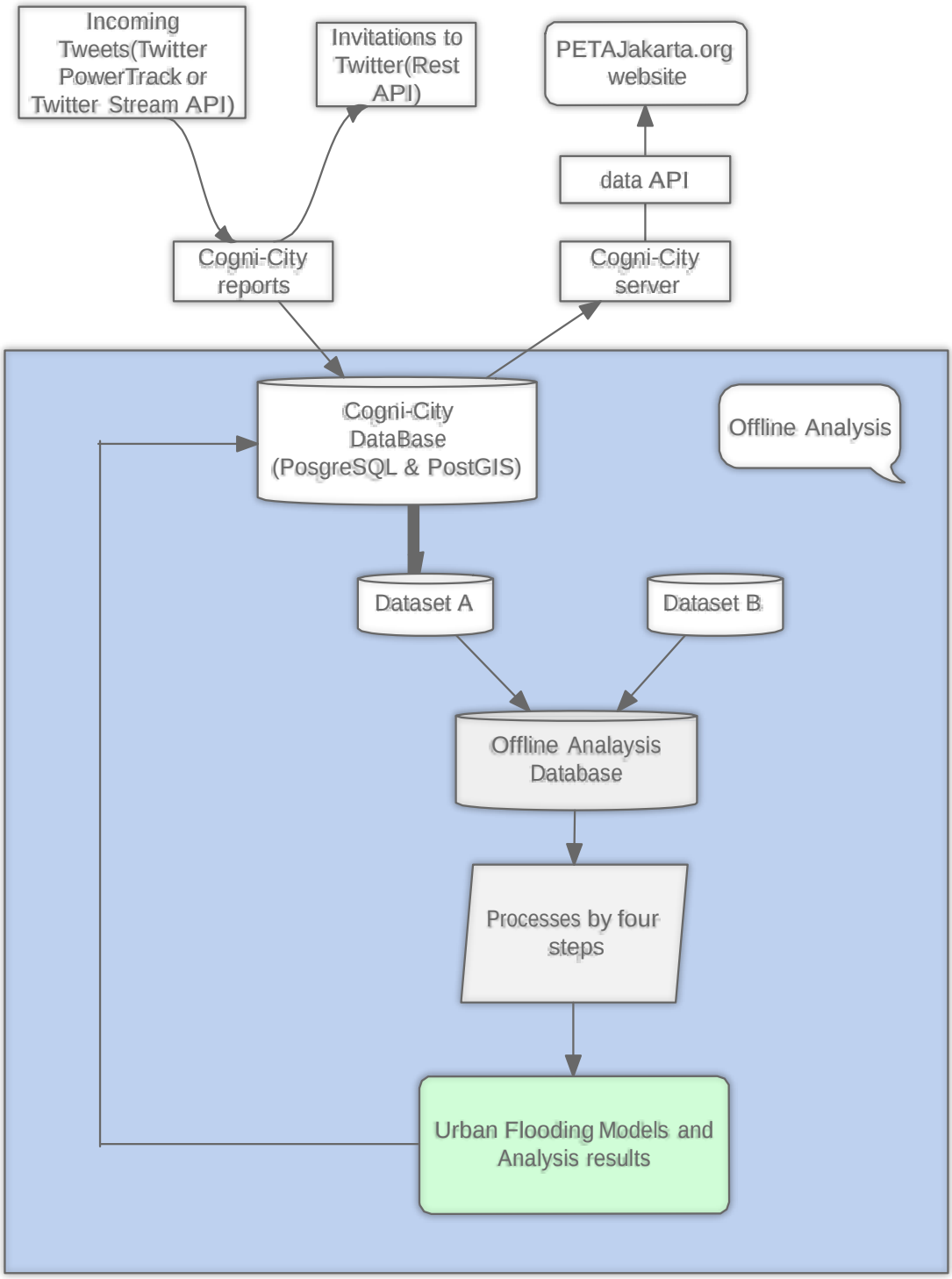

Figure 6. Research Diagram

The analysis methodology is divided into a series of data experiments that will use statistical tests to examine different aspects of the data. The four steps would be executed with four experiments that:

a) Experiment 1 - Summary statistics, which is to provide an overview of the data.
b) Experiment 2 – Time series plot, which is to explore the relationship between twitter activity and flood events
c) Experiment 3 – Relationship between Twitter activity and flood event, which is to test whether there is a relationship between number of tweets and number of flooded areas over time.
d) Experiment 4 – Spatial distribution of Twitter activity and flood event, which is to test whether tweet locations are related to flood event locations.

3.4.3 Software

In this study, the identified objectives will be fulfilled by the implementation of three stages, each of which corresponds, and directly relates to, a distinct experiment.

Analysis will be conducted by writing a suite of analysis software. This software will use the Psycopg library to query data using SQL from the PostGIS database for each experiment. A number of Python packages including numpy, scipi and matplotlib will then be used to undertake the calculations and plotting required.

In terms of processing within the three experiments, multiple software and tools are to be used, allowing the same data will be displayed using different styles. A number of relevant applications are described below.

1. QGIS:

QGIS is a cross-platform and open source GIS (Geographic Information System) application that provides data viewing, editing, and analysis capabilities. In addition, it provides integration with other open source GIS packages, such as PostGIS.

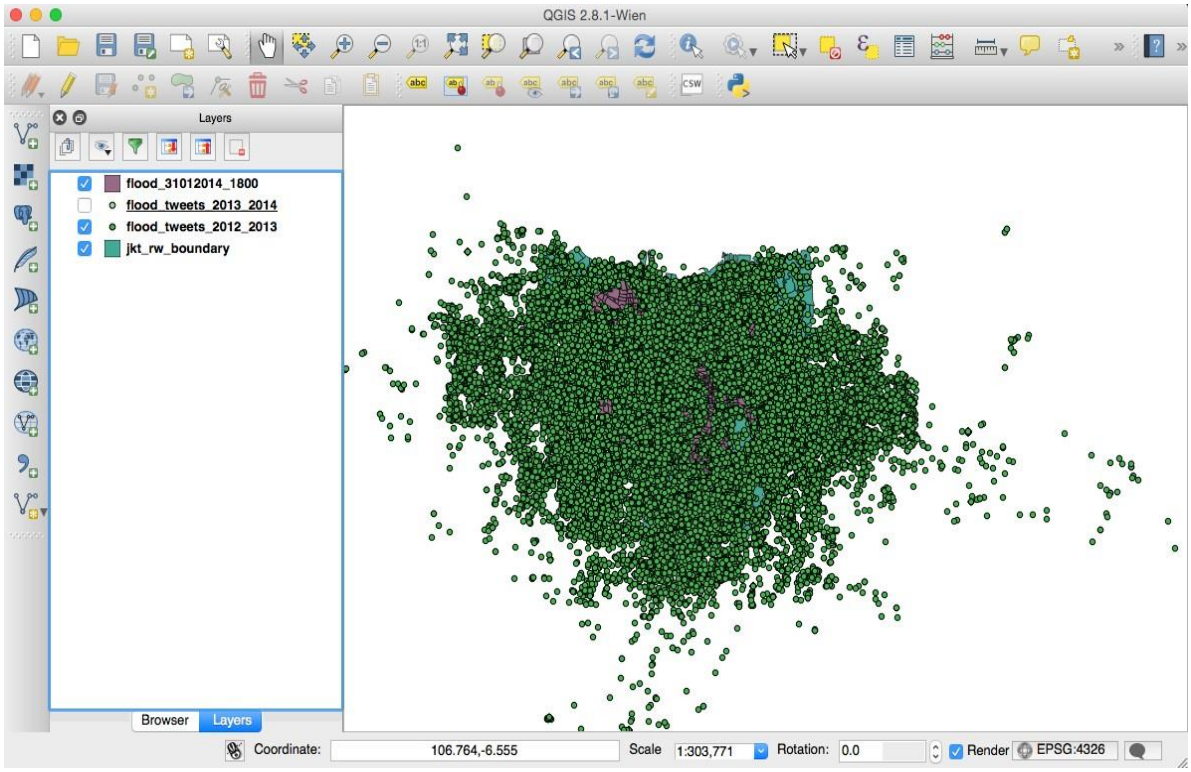

Figure 7. QGIS Screenshot

Figure 7 above displays the tweet data and the flooded areas of Jakarta on a digital map using the QGIS software. The brown polygons represent the areas that are affected at a certain time by flooding, which are supplemented with attributes such as coordinates and time. The green points represent all tweets during the monsoon period, and their respective geographic locations on the map. In addition, all of tweet points and flooded areas are shown at a same spatial scale, but different time scale.

2. pgAdmin 3:

The pgAdmin package is a feature-rich open source administration and development platform for PostgreSQL, and an advanced open source database management tool. PostgreSQL is an object-relational database management system (ORDBMS) with particular extensibility and standards-compliance, which implements the majority of the SQL.2011 standard. Figure 8 displays ‘Tables (81)’, in the left object browser, which refers to a collection of 81 tables of tweets data and corresponding flooded areas at particular points in time, whereby the flooded areas are recorded as ‘rw’.

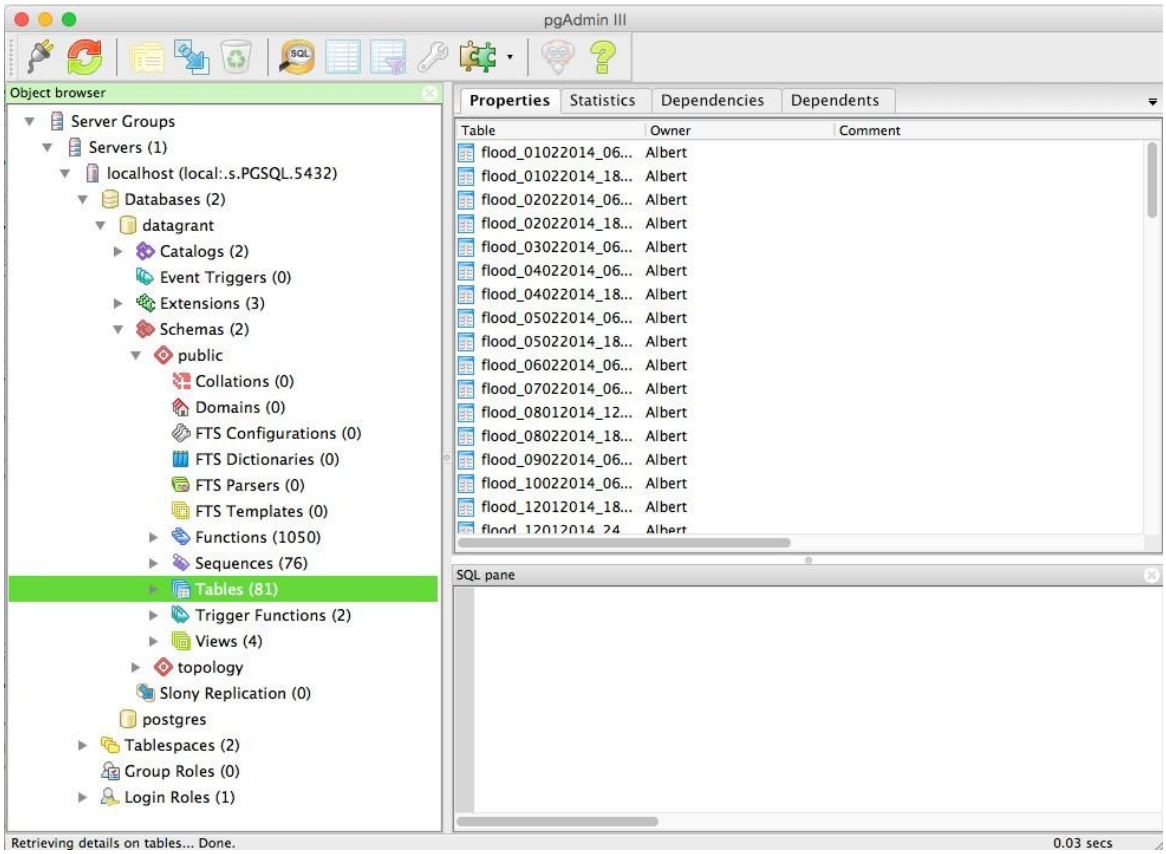

Figure 8. pgAdmin 3 Screenshot

3. Spyder 2:

Spyder is an open source cross-platform Integrated Development Environment (IDE) for scientific programming in the Python language, which includes NumPy, SciPy, Matplotlib and IPython integrations. It has the capability of connecting to the PostgreSQL database.

A fragment of data can be seen in Figure 9, in the right IPython console box. The data are printed as numbers with three columns: number of tweets, number of flooded areas, and time in the box, such as ‘50 0 2014-01-01 20:00:00’. For instance, the number ‘50’ shows there were 50 tweets at that time, ‘0’ indicates there were no areas affected by the flood, and the numbers to the right shows the specific time interval.

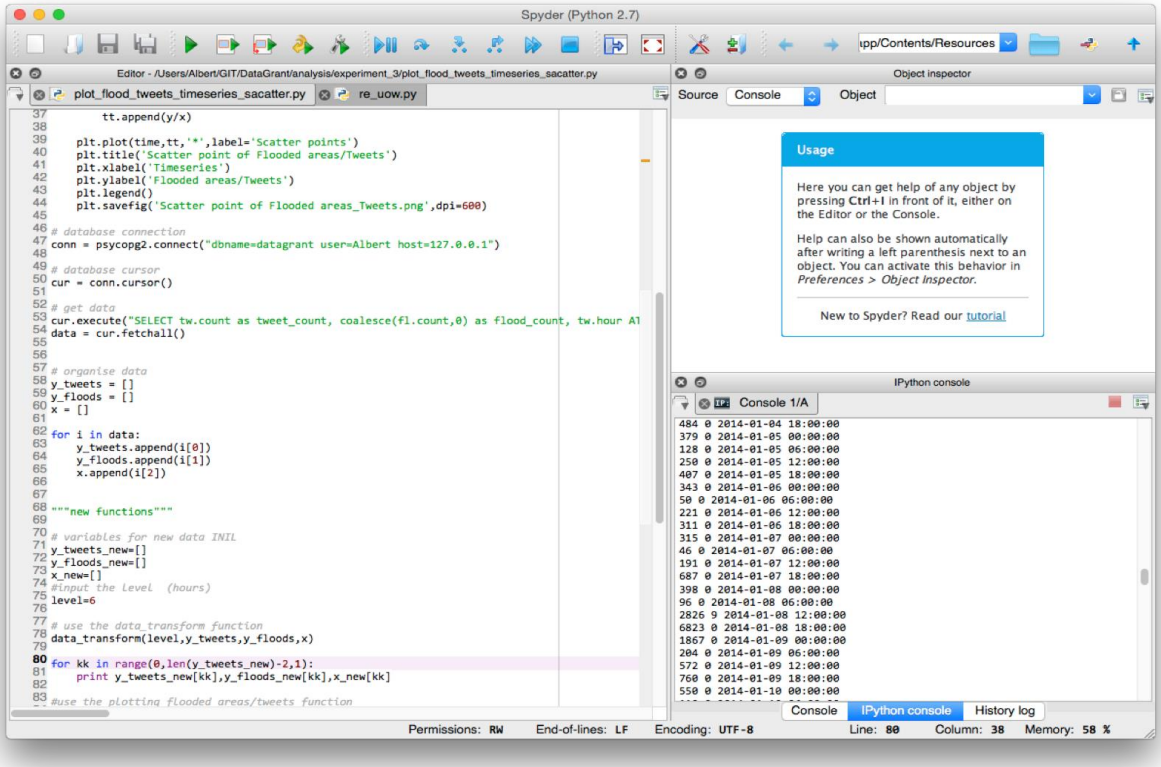

Figure 9. Spyder 2 Screenshot

4. SPSS:

SPSS (Statistical Product and Service Solutions) is a software package used for statistical analysis in social science. For the purpose of this research, SPSS is primarily used for comparative purposes, providing a comparison of the statistical analysis results from this package with that of the Python-based Spyder (see Figure 10). Specifically, SPSS was used to confirm that the results of both the Spearman Rank Correlation Coefficient, and the Linear Regression, are consistent.

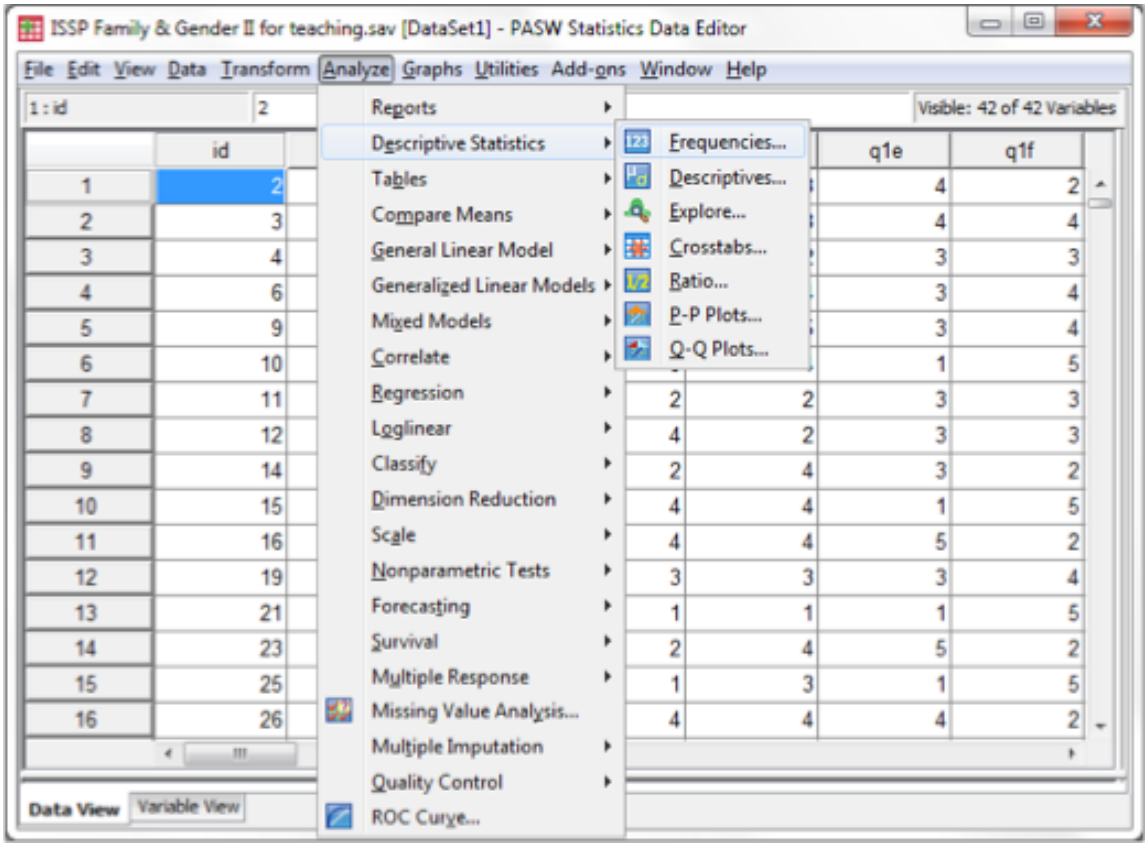

Figure 10. SPSS Screenshot

3.5 Experiments

The analysis methodology is divided into a series of stages that will use statistical tests to examine different aspects of the data. These stages are detailed below in terms of the specific calculations or series of tasks.

3.5.1 Summary Statistics

1. Objective: To provide an overview of the data
2. Twitter data:
    - Calculate the number of tweets for both monsoon seasons
    - Calculate the number of tweets with geo-location for both monsoon seasons
    - Calculate the number of original users both monsoon season
    - Plot a histogram of the number of tweets per user for each monsoon

3. Flood data:
    - Calculate the number of flood events (flooded areas) for each monsoon

3.5.2 Time Series Plot

1. Objective: To explore the relationship between Twitter activity and flood events
2. Twitter and flood data:
    - Plot the number of tweets in 12 hours over a time series
    - On the same plot (time series), add the number of flooded areas (RWs) per 12 hours

3.5.3 Relationship between Twitter Activity and Flood Events

1. Objective: To test whether there is a relationship between number of tweets and number of flooded areas over time
2. Null and Alternative Hypotheses:
    a. H_0: There will be no statistically significant relationship between the number of flooded areas and the number of tweets at the 95% confidence level.
    b. H_1: There will be a statistically significant relationship between the number of flooded areas (independent variable) and the number of tweets (dependent variable) at the 95% confidence level.
3. Twitter and flood data:
    - Calculate the number of flooded areas and the number of tweets at six-hourly intervals for both of the time series.
    - Create a scatter plot of the number of flooded areas/number of tweets
    - Calculate the Spearman's Rank Correlation Coefficient for these data with the number of tweets as the dependent variable at the 95% confidence level [48] .
    - If the trend appears linear, calculate simple linear regression for the relationship[48].

3.6 Data Analysis

The data analysis techniques heavily relied upon statistical analysis, including descriptive statistics, time series representations, clustering techniques, and data correlations. Tabular results, histograms, and scatter plots were specifically employed to show patterns and trends that might gleam specific insight into links between the number of tweets with the #banjir and declared flood management zones. The tweet was the unit of analysis at which this study was conducted, including the metadata of that tweet which might have also included location x and y coordinates and time stamps. Outside the scope of this thesis were the confirmed mapping of tweets on a visual GIS representation that could be used on a dashboard at am emergency operation centre to make decisions.

3.7 Conclusion

This chapter laid out the experiments that are to be conducted with respect to the PetaJakarta project. One of the complexities of the study was to not only extract the data from Twitter, but then to prepare the data for analysis. Additionally, how to take so many tweets and deduce significant insight from them was another challenge. The systems configuration of Cognicity dictated that a number of different software products were necessary for the experiments. This in itself was a significant integration process of which various skills were gained in order to master the capability. It was not only a means of using an existing statistical package to conduct the experiments, but to be responsible for the design and coding of the statistical analyses that would be reproduced with every dataset that needed to be analysed by clients.

Chapter 4. Results

4.1 Experiment One

4.1.1 Processes Employed to Achieve Objective 1

Based on the data from the Twitter users in Jakarta, it can be seen that not all tweets contained geo-location metadata. To determine the different aspects of data, PostgreSQL was used to count the numbers, shown as Figure 11 (a) and Figure 11 (b) below.

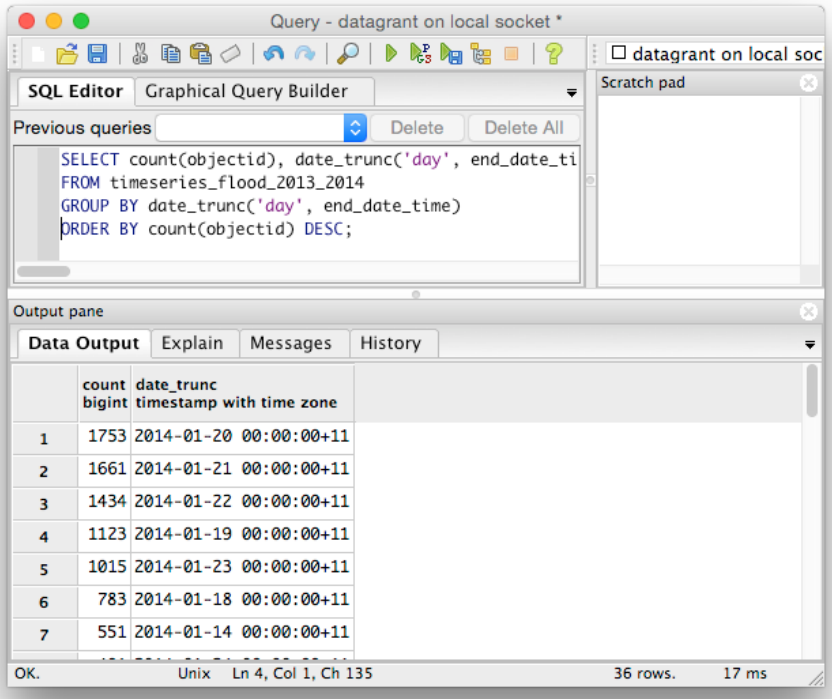

Figure 11. (a) SQL Selection

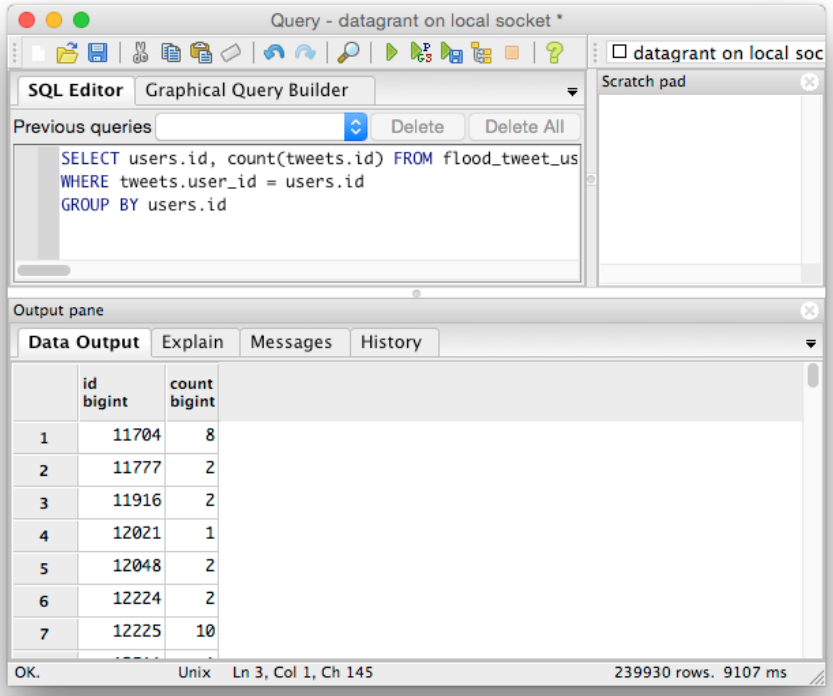

Figure 11. (b) SQL Selection

Please note that complete SQL details are provided in Appendix A, which contains fragments of the PostgreSQL and Python Codes used in this experiment.

To determine the density or frequency of tweets per Twitter user, a process of plotting took place using Spyder 2, based on the programming language Python. Figure 12 provides the programing interface screenshot.

tweets per users.py* | Spearman_rank_correlation（NEW）.py | re_uow.py

```
from pylab import *
# CONSTANTS
TBL_TWEETS = 'flood_tweets_2013_2014'
TBL_USERS = 'flood_tweet_users_2013_2014'
# database connection
conn = psycopg2.connect("dbname=datagrant user=Albert host=127.0.0.1")
# database cursor
cur = conn.cursor()
cur.execute("SELECT users.id, count(tweets.id) FROM flood_tweet_users_2012_2013 as users, flo
data = cur.fetchall()
cur.execute("SELECT users.id, count(tweets.id) FROM flood_tweet_users_2013_2014 as users, flo
data2 = cur.fetchall()
cur.close()

#savetxt("/home/albert/Data/albert_outcomes/per_users.csv",data,delimiter=",",fmt="%0.0f")
count = 0
peruser = []
for i in data:
    peruser.append(i[1])
    if count<i[1]:
       count=i[1]
    continue
count2 = 0
peruser2 = []
for i in data2:
    peruser2.append(i[1])
    if count2<i[1]:
       count2=i[1]
    continue

hist2=hist(peruser2,count2,facecolor="green",alpha=0.5)
plt.title('Number of tweets per user,2013/2014 Monsoon')
plt.xlabel('Number of tweets per user')
plt.ylabel('Count')
#plot(bins, count)
print count,count2
#savetxt("/home/albert/Data/albert_outcomes/peruser.csv",peruser,fmt="%0.0f")
plt.show()
cur.close()
conn.close()
```

Permissions: RW End-of-lines: LF En

Figure 12. Programing of tweets per user screenshot

Three packages were used in this instance: psycopg2, matplotlib, and numpy. The psycopg2 package was used for database processing, which can connect Spyder to the database and process the data stream by Python programming. The numpy package was used for mathematical processing. In all programs used for this research, the package matplotlib played a significant role in plotting all graphs. In this stage, the histogram generation function was employed to plot the density distribution or times of tweets per user (refer to the complete Python codes in Appendix A).

4.1.2 Experiment Outcomes

Table 2 shows the count of all tweets, tweets with geo-location metadata and corresponding Twitter users for each monsoon season. It additionally displays the number of flood events as measured by government data for the 2013-2014 monsoon. Please note that government flood event data was not available for the 2012-2013 monsoon.

Summary				
Monsoon	All tweets	With Geo-location	Original users	Flood events
2012-2013	1324853	71973	239930	N/A
2013-2014	1127069	167878	247989	658

Table 2 Summary Statistics

As the table shows, there were 239,930 original Twitter users but 1,324,853 tweets were tweeted out in the monsoon season 2012-2013. For the 2013-2014 period, there were 247,989 users and 1,127,069 tweets. This indicates that a large number of Twitter users tweeted more than once in both monsoon periods. To further illustrate this point, Figures 13(a) and 13(b) provide histograms for the two monsoons in order to show the density distributions or counts of tweets per user.

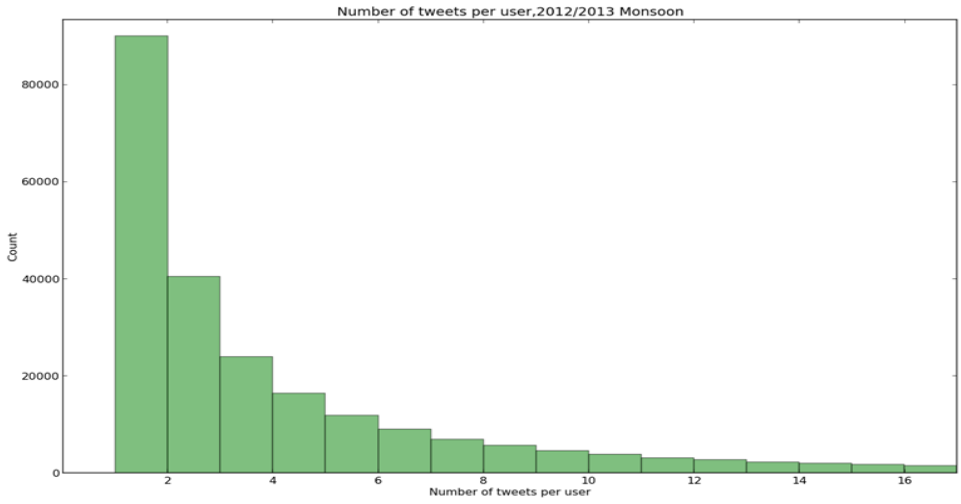

Figure 13. (a) tweets per user-2012-2013 monsoon

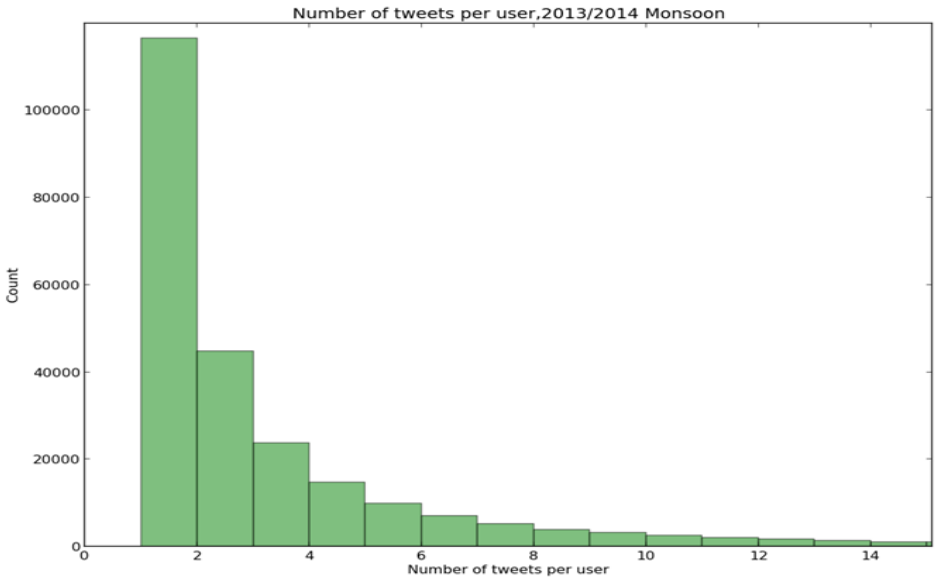

Figure 13. (b) tweets per user-2013-2014 monsoon.

The consistency, in terms of tweets per user, for both monsoon periods is evident. The two figures demonstrate that the majority of users tweeted once during flooding, about 90000 in monsoon 2012-2013 and 110000 in monsoon 2013-2014. Similarly, both histograms illustrate that users who tweeted twice during flooding in Jakarta are second in terms of count, and those who tweeted three times are third in terms of count, and so on. In summary, the graphs show that the counts of tweets per user tend to decline with the growth of the number of tweets per user. This evidence can explain why there were 239,930 original Twitter users in the monsoon period 2012-2013, whereas 1,324,853 tweets were tweeted. Likewise, for the monsoon period 2013-2014, there were 247,989 users while the tweets count was 1,127,069.

4.2 Experiment Two

4.2.1 Processes Employed to Achieve Objective 2

In this stage, the relationship between Twitter activity and flood events was explored by plotting the numbers of both Twitter activity and of flood events over the time series of the 2013-2014 monsoon season. As mentioned previously, formal government data on locations of flooding for the 2012-2013 monsoon was not available for comparison.

plot_flood_tweets_timeseries(any level).py | re_uow.py

```
import psycopg2 # database
import matplotlib.pyplot as plt # graphs
from numpy import *

#define transform function for different-hour-level
def data_transform(hour_level,orig_tweets,orig_floods,orig_time):
    loop_nb=0
    temp_ytw=0
    temp_yfl=0

    for j in range(0,len(orig_time)-1,1):
        if (x[j]-x[0+loop_nb*hour_level]).seconds<3600*hour_level:
            temp_ytw=temp_ytw+orig_tweets[j]
            temp_yfl=temp_yfl+orig_floods[j+1]
        else:
            y_tweets_new.append(temp_ytw)
            y_floods_new.append(temp_yfl)
            x_new.append(x[j])
            loop_nb=loop_nb+1
            temp_ytw=orig_tweets[j]
            temp_yfl=orig_floods[j+1]
    return;


# database connection
conn = psycopg2.connect("dbname=datagrant user=Albert host=127.0.0.1")

# database cursor
cur = conn.cursor()

# get data
cur.execute("SELECT tw.count as tweet_count, coalesce(fl.count,0) as flood_count, tw.hour Al
data = cur.fetchall()


# organise data
y_tweets = []
y_floods = []
x = []
```

Permissions: RW End-of-lines: LF

Figure 14. Programing of plotting screenshot

The fluctuations in the tweets and flood events over time are displayed visually on the same plot, so as to determine if there is a relationship between these two fluctuations. Counts of both Twitter activity and flood events are calculated at 12-hourly intervals. In the developed Python program, shown in Figure 14 (above), different hour-level (1 to 24) is available for plotting. The complete Python codes for Experiment Two are available in Appendix B. An algorithm was designed using the "data_transform" function to transform original data with 1 hour time interval to the data with 1~24 hourly time interval, as shown in Figure 14.

When the program is initiated by the user, the hour-level will be required as input for setting the time interval, as demonstrated in the console box in Figure 15.

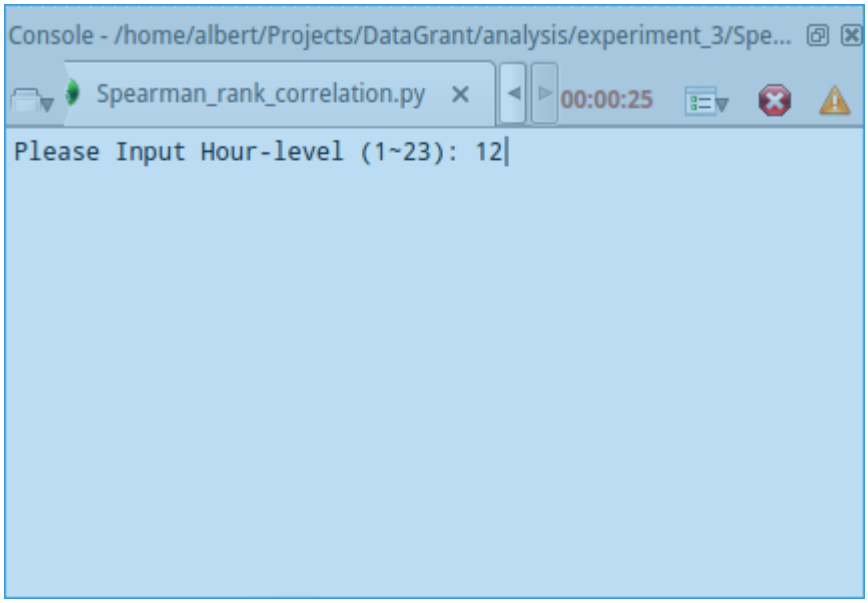

Figure 15. Spyder Console Screenshot.

4.2.2 Experiment Outcomes

In Figure 16, the green line represents Twitter activity which refers to the number of tweets, while the blue bars represent the flood events indicating the number of flooded areas.

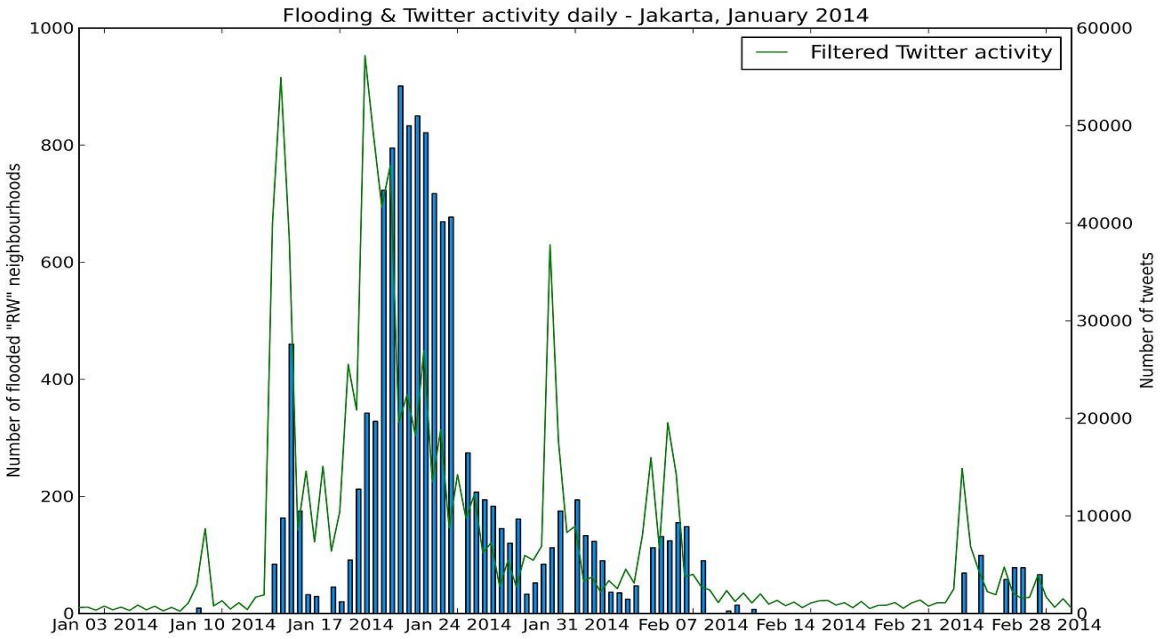

Figure 16. Plotting in 24-hours-level

The figure demonstrates that the fluctuation in Twitter activity is similar to changes in the number of flood events. Hence, there is an evident correlation between these two variables. Stage three, or experiment three will determine whether there is in fact a statistical correlation.

Additionally, as Figure 16 shows, the fluctuation is not always synchronous with the change of the flood events, which will be analysed in greater depth and comprehensively in the analysis provided further on in this chapter.

4.3 Experiment Three

4.3.1 Processes Employed to Achieve Objective 3

The purpose of this experiment was to test whether there is a statistical relationship between the number of tweets and the number of flooded areas over time. Consequently, the null hypothesis is that there will be no statistically significant relationship between the number of flooded areas and number of tweets at the 95% confidence level, and the alternative hypothesis is that there will be a statistically significant relationship between the number of flooded areas (independent variable) and the number of tweets (dependent variable) at 95% confidence level.

The Spearman Rank Correlation coefficient plays a significant role in geographic information systems (GIS) [48], since it is a nonparametric measure of the relationship between two sets of ordinal (ranked) values, which will be applied in this relationship test between the tweets and flood areas. The equation for the Spearman Rank Correlation coefficient is:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}.$$

Where ρ (i.e. r_s) is the Spearman rank correlation coefficient, d_i is the difference in ranking for each item which is: $d_i = x_i - y_i$, and n is the number of items ranked.

Therefore, this experiment required that the data be transformed into 6-hourly-level data to calculate the coefficient using a program developed in the Python programming language, an extract of which is shown in Figure 17, and complete details in Appendix C.

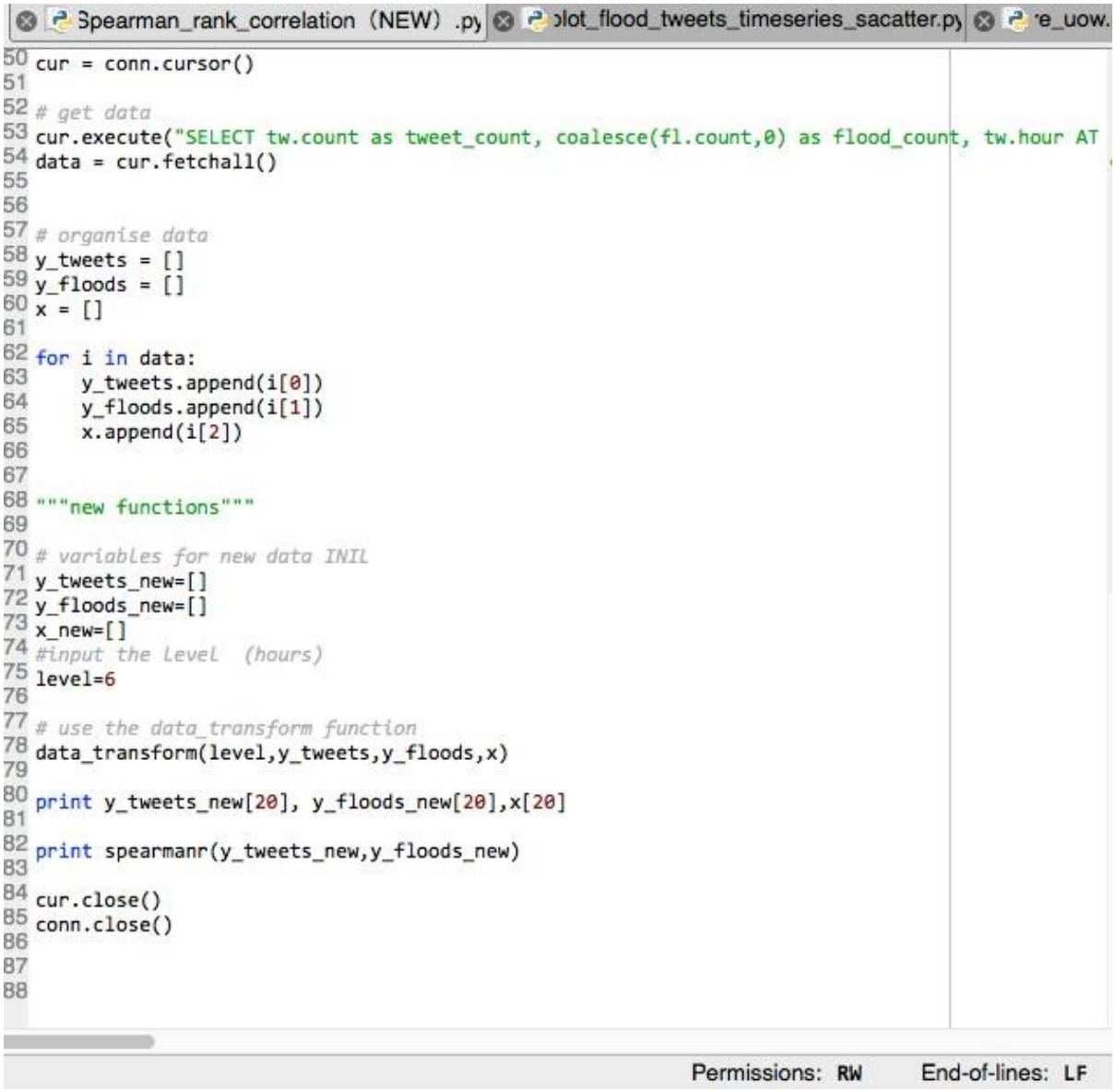

Figure 17. Programing of Spearman Coefficient screenshot

To accomplish the process of calculation, a module package of Python called “scipy” was used in this program, so as to calculate a Spearman rank-order correlation coefficient and to test for correlation by p-value. The Spearman correlation is a nonparametric measure of the monotonicity of the relationship between two datasets[48]. Unlike the Pearson correlation, the Spearman correlation does not assume that both datasets are normally distributed. Like other correlation coefficients, this one varies between -1 and +1 with 0 implying no correlation. Correlations of -1 or +1 imply an exact monotonic relationship. Positive correlations imply that as x increases, so does y. Negative correlations imply that as x increases, y decreases. The p-value roughly indicates the probability of an uncorrelated system producing datasets that has a Spearman correlation at least as extreme as the one

computed from the datasets. The p-values are not entirely reliable but are reasonable for datasets larger than 500 or so.

In the package "scipy", parameters are set as 'a' and 'b', which can be a 1-D or 2-D array, whereby 'b' is optional. Arrays contain multiple variables and observations. Each column of 'a' and 'b' represents a variable, and each row entry a single observation of those variables. Both arrays need to have the same length in the 'axis' dimension. Axis is either Integer or None, which is also optional.

If axis=0 (default), then each column represents a variable, with observations in the rows. If axis=0, the relationship is transposed: each row represents a variable, while the columns contain observations. If axis=None, then both arrays will be raveled.

At the end of this function, it returns the Spearman Rank Correlation Coefficient result rho: float or N-D array (2-D square). And the p-value, a float form and a same dimension as rho, is the two-sided hypothesis test whose null hypothesis is that two sets of data are not correlated. The Spearman correlation matrix or correlation coefficient (if only 2 variables are given as parameters) is square with length equal to total number of variables (columns or rows) in 'a' and 'b' combined.

To give a different visual representation of the correlation between tweets and flooded areas within the given time series, the scatter is determined by the matplotlib package. A screenshot of the program is provided in Figure 18, and full details of the corresponding Python codes can be found in Appendix C.

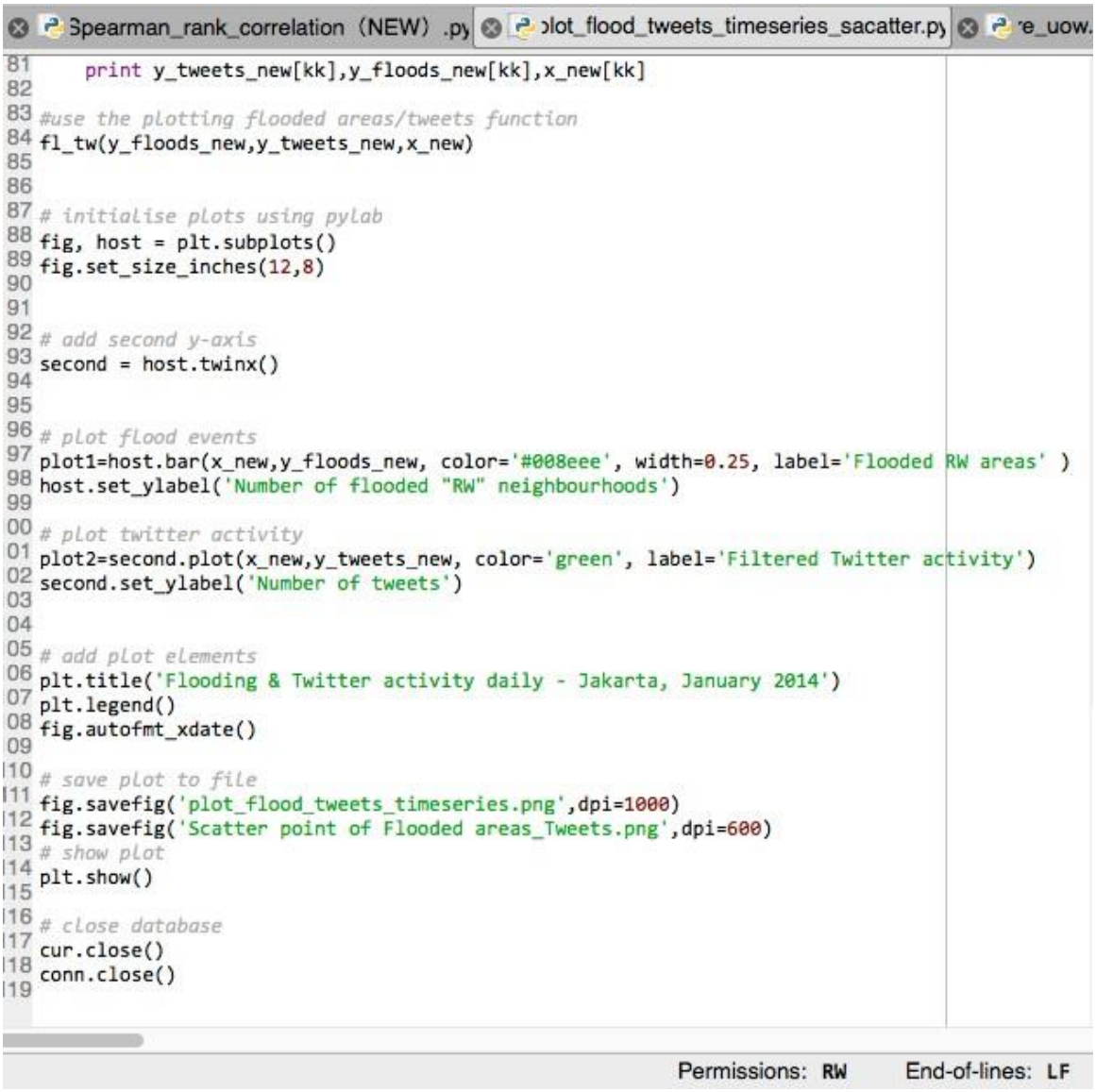

Figure 18. Programing time series scatter screenshot

4.3.2 Experiment Outcomes

A scatter plot of the number of flooded areas/number of tweets is created in Spyder by the Python package matplotlib, in order to provide a visual representation of the two variables, and test whether there is a linear regression between them visually at the first glance. As Figure 19 shows, the y-axis refers to the number of tweets and x-axis signifies the number of flooded areas. Every star represents the number of tweets per flooded area at a certain time interval. Figure 19 also demonstrates a significant number of stars on the y-axis, which means that these areas were not affected by flooding, while at the same time a large number of tweets are recorded. Those points refer to noise in this instance. Intuitively, if there is a strong

linear regression between these two variables, the scatter plot would reveal a linear trend. Although a strong linear regression cannot be seen graphically in Figure 19, this observation will be tested further using mathematical/statistical calculations in SPSS.

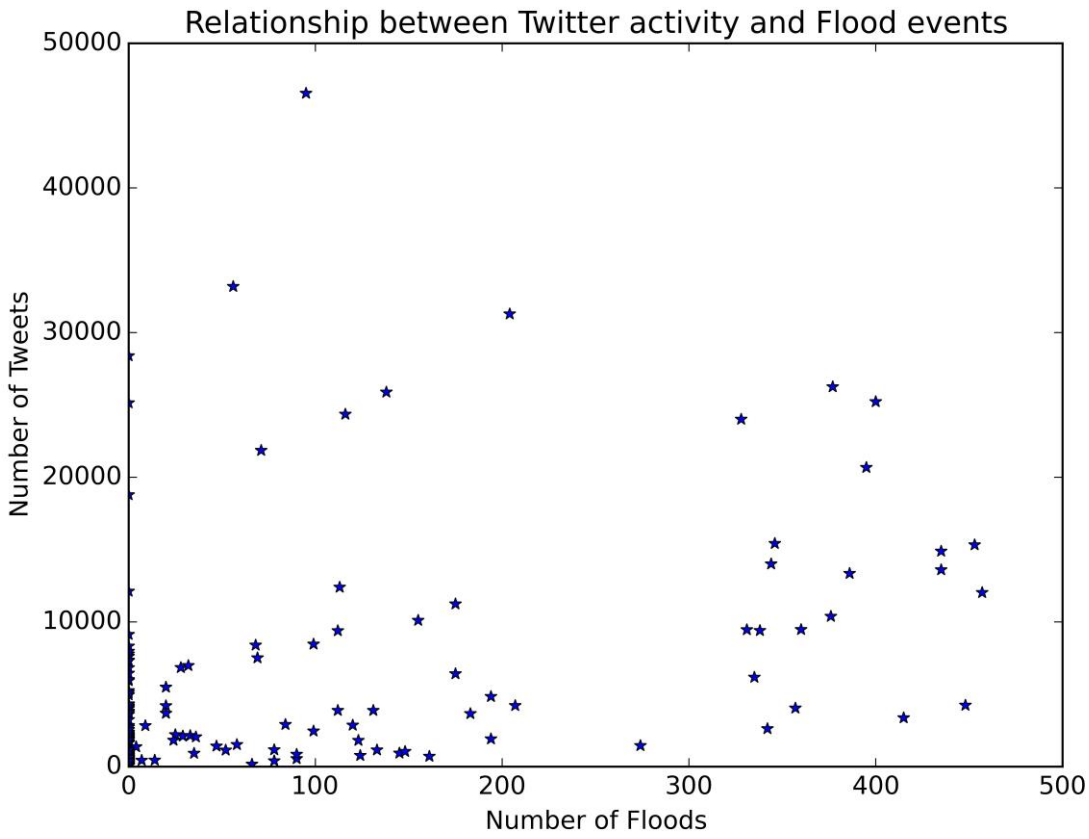

Figure 19. Scatter plot of the number of flooded areas/number of tweets

By using the Python package scipy, the result of Spearman Rank Correlation Coefficient rho was '0.54…' in the program, which is consistent with the result of rho as calculated by the SPSS package. Table 3 displays the results for the Spearman Rank Correlation Coefficient.

	Tools	Language	Equation	Method	Result
①	Spyder	Python	$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}.$	Scipy package that mentioned above	$\rho(r_s)$ = 0.54 reject H_0 accept H_1
②	SPSS (Statistical Product and Service Solutions)	SPSS	$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}.$	functions nested in SPSS	$\rho(r_s)$ = 0.54 reject H_0 accept H_1

Table 3. Spearman Rank Correlation Coefficient Results

Significantly, the two calculation processes of the Spearman Rank Correlation Coefficient are at a 95% confidence level. Therefore, based on this result, the null Hypothesis H_0 is rejected: that there is no statistically significant relationship between the number of flooded areas and the number of tweets at the 95% confidence level, and the alternative Hypothesis H_1 is accepted. Furthermore, the linear regression of numbers of flooded areas and tweets was tested using SPSS. The results related to significance (P-value), a and b are shown in Table 4.

N	Sig. (P value)	A	b	X	Y
239	2.227E-15	2358.82	28.73	Flooding events	Tweets

Sig. (P value) : 2.227E-15<<0.05 which can prove that the established regression equation of statistical significance, namely the linear relationship between independent variable and dependent variable

Table 4. Linear Regression outcomes

Based on the results in Table 4, the linear regression equation of tweets and flooded areas is presented:

Linear regression equation: Y=2358.82 + 28.73*X

In addition, to provide a visual representation of the linear regression, the linear trend based on the equation in SPSS is shown as Figure 20.

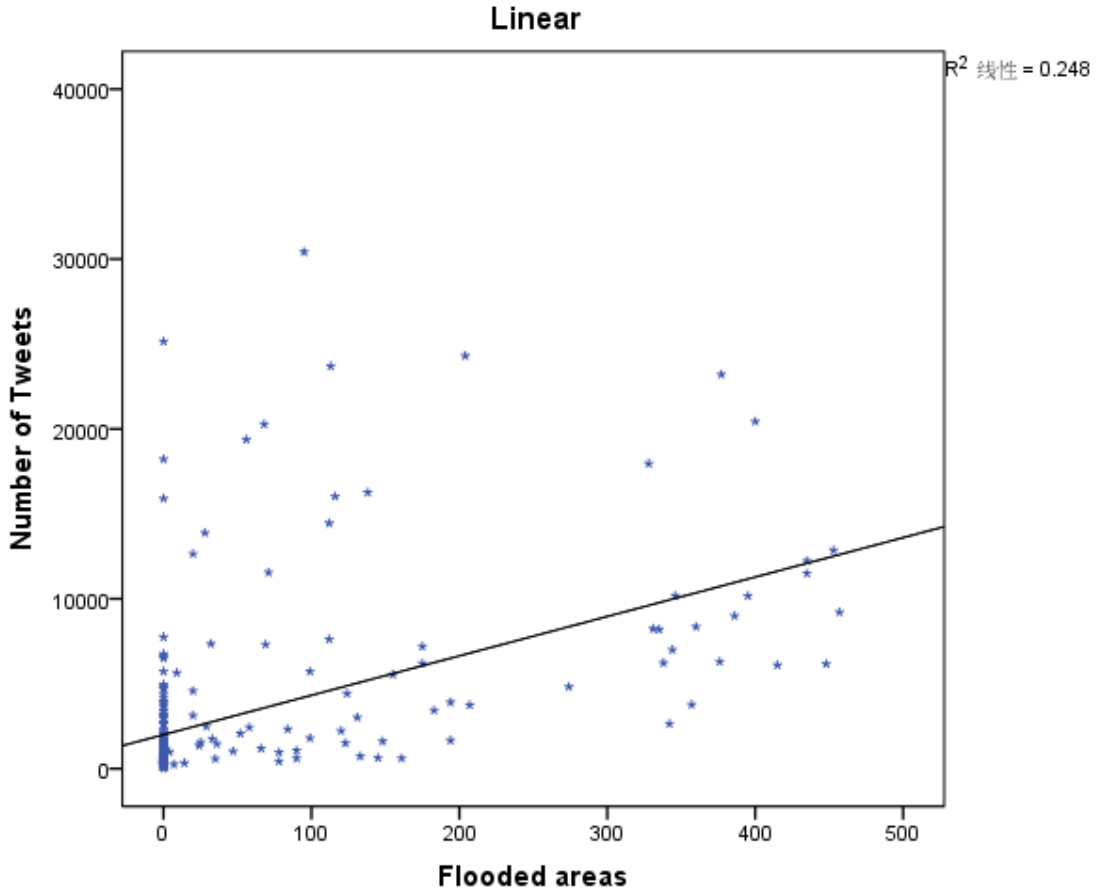

Figure 20. Linear regression scatter and line

4.4 Discussion and Analysis

The purpose of this section is to analyse the relationship between the flood events and Twitter activity using the data itself in addition to the real-world situations, so as to provide a greater understanding of the relationship between the flooded areas and tweets based on the results of the experiments presented above. The analysis process involves dividing the six-hourly histogram into 3 parts, labelled as '①' to '③', by the two yellow lines shown in Figure 21, which symbolise the preliminary, middle and end phases of flooding events.

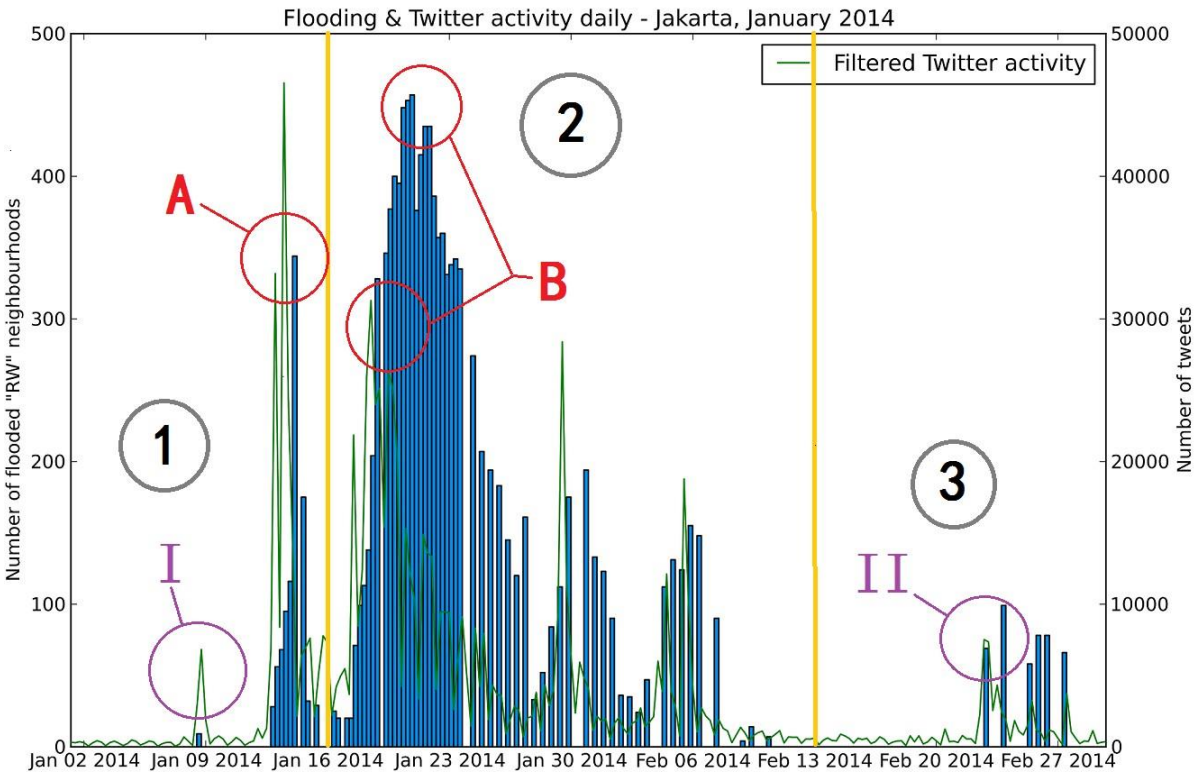

Figure 21. Analysis Illustration of relation between Flooding & tweets

4.4.1 Findings and Illustrations

Figure 21 illustrates that the general fluctuation of tweets is not always synchronous with the change of the flood events except at a few time intervals. The highest number of tweets, close to 48000, occurred in the preliminary phase instead of the middle phase when the greatest number of flood events had occurred, peaking at 470, as recorded in the government data. Additionally, it can be seen, when referring to the two sites marked in red ‘A’ and ‘B’, that the number of tweets declines at the same point at which the flooded areas increase.

To fully understand the phenomenon mentioned above, it is helpful to compare with evidence of observational data from the real situation in Jakarta. For instance, at the time of the flooding in Jakarta:

i. People are unable to utilise mobile phones given the need to flee the severely flooded sites.

ii. Due to power failures, mobile phones cannot be charged in certain areas where civic infrastructures are affected by flooding.

4.4.2 Benefits of Twitter in Floods

Based on these findings, Twitter activity, notably the intelligence gained from individuals responding to flood events using this medium, provides a number of benefits.

1) Critical value

Citizens' Twitter activity relating to floods is dependent on the flooding events. One would generally expect that the more severe the flooding events, the more intensive or frequent the Twitter activity. The observed trend in this study, however, is that while there was an increased response at the beginning of every monsoon period, the trend altered and Twitter activity decreased with the intensification of the flooding, as is demonstrated by the red circles corresponding to the letter 'B' in the Figure 19. That is, Twitter activity decreased intensively while the flooded events increased sharply. This is similarly the case for flooding events marked by the letter 'A' in Figure 19, whereby tweets decreased with the increase of flooding events. Therefore, the critical scale of flooding which can change the trend is presented here based on the reality. This means that the Twitter activity will decrease at a certain scale, which is termed 'critical value', of flooding with the growth of the scale of flooding. It is reasonable to state that citizens will forgo tweeting when the flooding scale is too high over the critical value. In this case, the number of tweets declined sharply.

Referring to the data, two variables are of importance: the number of tweets (six-hourly) which is the dependent variable, and number of flooding events (six-hourly) which is the independent variable. Table 5 (a) and (b) offers fragments of data (six-hourly) in which the left column signifies the numbers of tweets and the right column records the numbers of flood events.

44	1271	0
45	608	0
46	1293	0
47	6839	28
48	33184	56
49	8386	68
50	46545	95
51	24351	116
52	14001	344
53	2160	0
54	6409	175
55	6974	32
56	7614	0
57	2108	29

Table 5(a)

63	4953	0
64	5483	20
65	3680	20
66	21857	71
67	8466	99
68	12400	113
69	25878	138
70	31290	204
71	24002	328
72	25134	0
73	15417	346
74	26248	377
75	25224	400
76	20675	395
77	4240	448
78	15323	453
79	12031	457
80	10387	376
81	3363	415
82	14888	435
83	13600	435
84	13349	386

Table 5(b)

In Table 5(a), the highlighted fields indicate that the number of tweets increased to 46,545 when the number of flooding events was 95 (and increasing). After the peak point in flooding events (344), the number of tweets decreased dramatically to 14,001. Comparably, in Table 5(b), the number of tweets increased to the highest point of 31,290, after which the figures declined with the increase in flooding events (328). Furthermore, when the number of flooded areas reached a peak of 457, the number of tweets dropped to 12,031.

While this does not provide the accurate critical value, it does illustrate the real-world situation, explaining the unexpected points that make the fluctuation of the tweets asynchronous with changes in the flood events.

2) Tolerance

In the beginning phase of flooding in the monsoon period of 2013-2014, citizens in Jakarta who were suffering as a result of the events were highly reactive to the disaster. This is demonstrated by comparing and contrasting the preliminary and end phases of the monsoon. This involves an analysis of peak numbers of flooded areas and tweets around certain time intervals, shown in Figure 22 (a) and Figure 22 (b).

It can be seen in the preliminary phase in Figure 22(a) that there were approximately 6800 tweets from about 9 flooded areas at a particular point in time. The initial flooding event is marked by the letter 'I' (Figure 21). The number of tweets is 700 times greater than the number of flooded areas. When the number of tweets reached 33000, the number of flooded areas also increased to 60, and similarly the amount of tweets was still much greater (550 times) than the number of flooded areas at that exact point in time.

Nevertheless, in the end phase, there had not been any flooding events for a series of days. Shortly after, when a flooding event occurred, the number of tweets increased to the highest point of about 7,500 corresponding to about 70 flooded areas. The number of tweets was approximately 100 times greater than flooded areas around the same time period.

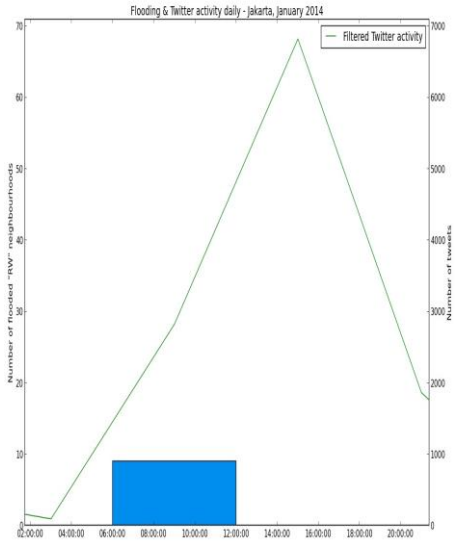

Figure 22(a)

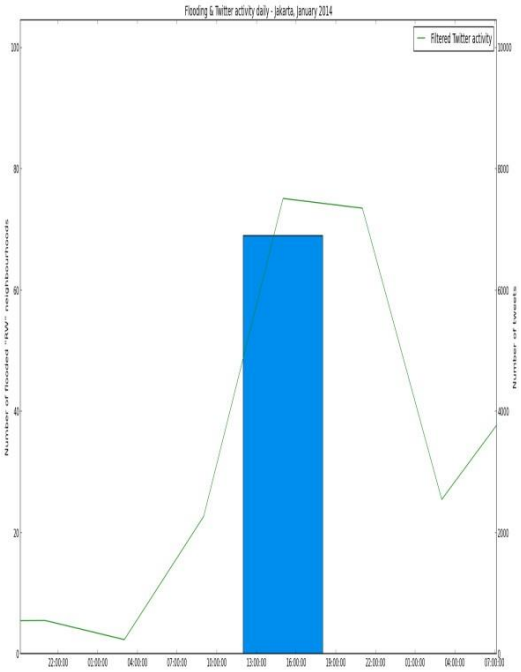

Figure 22(b)

Based on the findings showing that the numbers differ between the preliminary and final phases, it can be concluded that the reactivity of citizens to flooding decreased from 700 and 550 to 100. Therefore, the term 'tolerance' is proposed to explain this phenomenon.

Chapter 5. Conclusion

5.1 Introduction

This chapter provides a summary of the principal findings of this thesis, its major contributions to the disaster management domain and the importance of geosocial intelligence in this field. It enhanced the capacity to understand and promote the resilience of cities to both extreme weather events due to climate change and to long-term infrastructure transformation with the process of climate adaptation, and contribute to a wider understanding of the potential of social media to act as a real time crowd-sourcing tool during extreme weather events.

The limitations of the study are discussed: spatial data mining technology, when experts want to adjust to recent data sorts such as uncertain and sparse geo-referenced social media feeds. Also future research directions are suggested: quantifying the utility of social media data during flood events within the context of a civic co-management framework.

5.2 Principal Findings and Major Contributions

As a part of and based on the project PETAJakarta.org, this research has been conducted using a systematic approach, and in an objective manner, whereby the results are derived primarily from crowd-sourced data from everyday citizens. It demonstrates the relationship between Twitter activity and flood events using statistical means. Past Twitter data and the real observed flooding events are used as the basis for modelling the urban flooding event in its totality. What is more, the increasing utilization of computer linguistic approaches with associated spatiotemporal analysis to deduce textual messages from Twitter, becomes an emerging field which is currently still lacking real-world implementation beyond small scoped studies. In addition, new doors for the investigation in the field of GIScience were opened combined GIS with location-based social networks. Without a doubt, with the methods offered by GIScience, the spatiotemporal extraction and analysis of real-time social- media information, such as data from Twitter, could be much more convenient and effective.

The 4 principal findings of the study included:

1. Most Twitter users tweet once during flooding using keywords in their native language.

This indicates that a large number of Twitter users tweeted more than once in both monsoon periods. That means the counts of tweets per user tend to decline with the growth of the number of tweets per user.

2. The fluctuation in the number of tweets is not consistent with the number of flooded areas.

   The figures of experiment 2 demonstrate that the fluctuation in Twitter activity is similar to changes in the number of flood events. Therefore, there is an evident correlation between these two variables. And the fluctuation is not always synchronous with the change of the flood events, which was analyzed in greater depth and comprehensively in the analysis.

3. There is a statistically significant relationship between the number of flooded areas and the number of tweets at the 95% confidence level.

   A scatter plot of the number of flooded areas/number of tweets is created in Spyder by the Python package matplotlib, in order to provide a visual representation of the two variables, and test whether there is a linear regression between them visually at the first glance. And the same observation was tested using mathematical/statistical calculations in SPSS.

   Significantly, the two calculation processes of the Spearman Rank Correlation Coefficient are at a 95% confidence level. Based on this result, the null Hypothesis H_0 is rejected: that there is no statistically significant relationship between the number of flooded areas and the number of tweets at the 95% confidence level, and the alternative Hypothesis H_1 is accepted.

4. There is a linear regression between the number of flooded areas and the number of tweets.

   And, the linear regression of numbers of flooded areas and tweets was tested using SPSS. Based on the results, the linear regression equation of tweets and flooded areas is presented:

   Linear regression equation: $Y=2358.82 + 28.73*X$

5. Comparing with evidence of observational data from the real situation in Jakarta, here comes to 3 suppositions:
    a. People are unable to utilise mobile phones given the need to flee the severely flooded sites.
    b. Due to power failures, mobile phones cannot be charged in certain areas where civic infrastructures are affected by flooding.
    c. Citizens' Twitter activity relating to floods is dependent on the flooding events. The more severe the flooding events, the more intensive or frequent the Twitter activity. There was an increased response at the beginning of every monsoon period, the trend altered and Twitter activity decreased with the intensification of the flooding. The Twitter activity will decrease at a certain scale, which is termed 'critical value', of flooding with the growth of the scale of flooding. It is reasonable to state that citizens will forgo tweeting when the flooding scale is too high over the critical value. In this case, the number of tweets declined sharply. While this does not provide the accurate critical value, it does illustrate the real-world situation, explaining the unexpected points that make the fluctuation of the tweets asynchronous with changes in the flood events.

5.3 Limitations and Next Steps

Given the complex nature of flooding, this research did not develop effective mathematical models to demonstrate the relationship between the flooding events and Twitter activity. This was beyond the scope of the thesis.

Developing an understanding of the spatial and temporal relationships between Twitter activity and flooding at the "rw" level in Jakarta, with the aim of producing a trigger metric for flooding based on Twitter activity, requires further experimentation. It is suggested that such an experiment be based on the following three steps:

1. Examine the spatial distribution of Twitter activity across the city during flood events using spatial clustering analysis, and compare this qualitatively to spatial distribution of flooding.
2. Perform a statistical test that calculates the difference between the counts of spatiotemporally paired Twitter activity related to flooding and flooding events

3. Perform a statistical test on counts of tweets before and after flood events to see if there is a significant different between the flooding events and Twitter activity.

It is therefore suggested that a fourth experiment be conducted as an extension to this research, the details of which are as follows:

1. Objective: To test whether tweet locations are related to flood event locations
2. Null and Alternative Hypothesis:
    a. H_0: The number of tweets in flooded areas will be significantly greater than that of non-flooded areas at the 95% confidence level.
    b. H_1: There will be a greater number of tweets in flooded areas at 95% confidence level.
3. Twitter and flood data:
    a. For each monsoon calculate the total area flooded (i.e. November to March)
    b. Calculate the proportion of tweets expected to be within flooded areas. If 40% of the city was flooded, and a total of 1000 tweets were recorded, we would expect 400 tweets to be within flooded areas.
    c. Calculate the number of tweets inside flooded areas using the PostGIS point in polygon function.
    d. Perform the Chi-Squared test on the number of tweets in flooded areas and the number of tweets outside flooded areas to see if the proportion within the flooded area is significantly higher than the expected proportion[48].

5.4 Future Research

The number of papers now being published in the domain of geosocial intelligence is growing rapidly as researchers, industry and government recognise the potential to harness big data spatiotemporal techniques toward solving problems that affect large urban populations, particularly in developing nations. The media in particular have fully supported the initiatives with social impact publications in major outlets such as BBC, CCN, HuffingtonPost, National Geographic, ODI, and much more. Researchers from across disciplines are working together to solve global challenges that have come about from human-made interventions on existing physical geographies to the detriment of highly urbanised communities. The PetaJakarta project has now been expanded by some of the original members of the project team, to investigate other locations in Indonesia, see for

example, https://petabencana.id/. Part of the original team have also now begun putting their ingenuity toward the Urban Risk Lab at MIT http://urbanrisklab.org/.

Beyond emergency management, crowd-sourcing big data initiatives are now being considered toward large-scale environmental sustainability measures and humanitarian needs. Increasing the resilience of local communities is the primary aim of project work now being spearheaded throughout the world. While we cannot prove that geosocial intelligence can help us better respond to natural disasters, it may well be able to raise enough awareness to implement strategies that might mitigate them over the longer-term. At the very least aggregating geographical and social media datasets together helps to shed light on what happens during emergencies, how individuals in the community respond, what to expect, and how to better prepare at responding the next time such an event occurs. Applying a variety of analytical techniques on Twitter data sets can enrich our understanding even more, such as content analysis, sentiment analysis and even conversational analysis. Increasing resilience within the community should be another primary aim of future research. Of course, such techniques will lend themselves open to security attacks, how to ensure the datasets is free of noise or misinformation, denial of service attacks through bot messaging, and more. The introduction of social machines, and machine-to-machine communications through the Internet of Things may well act to exacerbate breaches and currently the only way to confirm a flood report from a civilian end-user is to automatically reply to their tweet appropriately.

Over time, it is also expected that more tweets will come with location metadata attached, not just time stamps which will allow for more accurate visual map representation. Users will also become more competent at communicating over short messages, the greater the penetration of smartphones and social media apps. But the integration of various autonomous data collection systems will additionally change the way planning occurs in cities, everything from the sharing of data collected via personal wearable devices, to sensors attached to smartgate systems (e.g. canals and other infrastructure). Geographical information systems have the power to bring together all the various datasets for interrogation by authorities tasked with planning and development.

5.5 Conclusion

This research was a trial study that aims to model urban flooding using geosocial intelligence, notably Twitter data supplemented with government-based information. The objectives of the

study have been successfully achieved, and further research directions demonstrating how the study can be further developed have been proposed. It summary, it is evident that the potential of social media and geosocial intelligence in the (natural) disaster management realm are great, but are still being explored and researched. This study provides a valuable contribution to this growing body of literature. If Twitter was launched in 2006, and smartphones began to proliferate around 2009, and PetaJakarta was officially launched in 2014, there is great scope for the realm of possibility. No doubt with about 2 billion smartphone users in the world today, the majority of whom are active social media users, the way we harness people as sensors will continue to mature. With that pervasiveness of personalised data collection will come an increased spotlight onto the ethical use of this information.

Appendices

Appendix A - Experiment 1

1. Codes for PostgreSQL:

a) SELECT count(objectid), date_trunc('day', end_date_time) FROM timeseries_flood_2013_2014 GROUP BY date_trunc('day', end_date_time) ORDER BY count(objectid) DESC;

b) SELECT users.id, count(tweets.id) FROM flood_tweet_users_2012_2013 as users, flood_tweets_2012_2013 as tweets WHERE tweets.user_id = users.id GROUP BY users.id

2. Codes for Python:

```
import os import

json import

psycopg2

from pylab import *

# CONSTANTS

TBL_TWEETS = 'flood_tweets_2013_2014'

TBL_USERS = 'flood_tweet_users_2013_2014'

# database connection

conn = psycopg2.connect("dbname=albert user=albert password=tale.007 host=127.0.0.1")

# database cursor

cur = conn.cursor()

cur.execute("SELECT users.id, count(tweets.id) FROM flood_tweet_users_2012_2013 as users, flood_tweets_2012_2013 as tweets WHERE tweets.user_id = users.id GROUP BY users.id")
```

```
data = cur.fetchall()

cur.execute("SELECT users.id, count(tweets.id) FROM flood_tweet_users_2013_2014 as users, flood_tweets_2013_2014 as tweets WHERE tweets.user_id = users.id GROUP BY users.id")

data2 = cur.fetchall()

cur.close()

#savetxt("/home/albert/Data/albert_outcomes/per_users.csv",data,delimiter=",",fmt="%0.0f")

count = 0

peruser = []

for i in data:

    peruser.append(i[1])

    if count<i[1]:

        count=i[1]

    continue


count2 = 0

peruser2 = []

for i in data2:

    peruser2.append(i[1])

    if count2<i[1]:

        count2=i[1]

   continue


hist2=hist(peruser2,count2,facecolor="green",alpha=0.5)

plt.title('Number of tweets per user,2013-2014 Monsoon')
```

```
plt.xlabel('Number of tweets per user')

plt.ylabel('Count')

#plot(bins, count)

print count,count2

#savetxt("/home/albert/Data/albert_outcomes/peruser.csv",peruser,fmt="%0.0f")

plt.show()

cur.close()

conn.close()
```

Appendix B - Experiment 2

1. Codes for PostgreSQL:

```
SELECT tw.count as tweet_count, coalesce(fl.count,0) as flood_count, tw.hour AT
TIME ZONE 'ICT'  FROM(SELECT count(id) as count, date_trunc('hour',
posted_time) as hour FROM flood_tweets_2013_2014 WHERE date_trunc('hour',
posted_time) AT TIME ZONE 'ICT' >= '2014-01-01' AND date_trunc('hour',
posted_time) AT TIME ZONE 'ICT' < '2014-03-02' GROUP BY hour) as tw LEFT
OUTER JOIN (SELECT count(objectid) as count, date_trunc('hour', end_date_time) as
hour FROM timeseries_flood_2013_2014  GROUP BY date_trunc('hour',
end_date_time) ) as fl   ON tw.hour = fl.hour ORDER BY tw.hour
```

2. Codes for Python:

```
import psycopg2 # database

import matplotlib.pyplot as plt # graphs

from numpy import *



#define transform function for different-hour-level

def data_transform(hour_level,orig_tweets,orig_floods,orig_time):

    loop_nb=0

    temp_ytw=0

    temp_yfl=0



    for j in range(0,len(orig_time)-1,1):

        if (x[j]-x[0+loop_nb*hour_level]).seconds<3600*hour_level:

            temp_ytw=temp_ytw+orig_tweets[j]

            temp_yfl=temp_yfl+orig_floods[j+1]
```

```
        else:
            y_tweets_new.append(temp_ytw)
            y_floods_new.append(temp_yfl)
            x_new.append(x[j])
            loop_nb=loop_nb+1
            temp_ytw=orig_tweets[j]
            temp_yfl=orig_floods[j+1]
    return;


# database connection
conn = psycopg2.connect("dbname=datagrant user=Albert host=127.0.0.1")


# database cursor
cur = conn.cursor()


# get data
cur.execute("SELECT tw.count as tweet_count, coalesce(fl.count,0) as flood_count, tw.hour AT TIME ZONE 'ICT' FROM(SELECT count(id) as count, date_trunc('hour', posted_time) as hour FROM flood_tweets_2013_2014 WHERE date_trunc('hour', posted_time) AT TIME ZONE 'ICT' >= '2014-01-01' AND date_trunc('hour', posted_time) AT TIME ZONE 'ICT' < '2014-03-02' GROUP BY hour) as tw LEFT OUTER JOIN (SELECT count(objectid) as count, date_trunc('hour', end_date_time) as hour FROM timeseries_flood_2013_2014 GROUP BY date_trunc('hour', end_date_time) ) as fl ON tw.hour = fl.hour ORDER BY tw.hour")
data = cur.fetchall()


# organise data
```

```
y_tweets  =  []

y_floods  =  []

x = []


for i in data:

    y_tweets.append(i[0])

    y_floods.append(i[1])

    x.append(i[2])


"""new functions"""

# variables for new data INIL

y_tweets_new=[]

y_floods_new=[]

x_new=[]


#input the level  (hours)

level=int(raw_input("please  input level(hourly):  "))

# use the data_transform  function

data_transform(level,y_tweets,y_floods,x)


# initialise plots using pylab

fig, host = plt.subplots()

fig.set_size_inches(12,8)
```

```
# add second y-axis

second = host.twinx()


# plot flood events

plot1=host.bar(x_new,y_floods_new,  color='#008eee', width=0.25, label='Flooded RW areas' )

host.set_ylabel('Number  of flooded "RW" neighbourhoods')


# plot Twitter activity

plot2=second.plot(x_new,y_tweets_new, color='green', label='Filtered Twitter activity')

second.set_ylabel('Number  of tweets')


# add plot elements

plt.title('Flooding  & Twitter activity daily - Jakarta, January 2014')

plt.legend()

fig.autofmt_xdate()


# save plot to file

fig.savefig('plot_flood_tweets_timeseries.png',dpi=600)


# show plot

plt.show()

# close database

cur.close()

conn.close()
```

Appendix C - Experiment 3

1. Codes for PostgreSQL:

```
SELECT tw.count as tweet_count, coalesce(fl.count,0) as flood_count, tw.hour AT
TIME ZONE 'ICT' FROM(SELECT count(id) as count, date_trunc('hour',
posted_time) as hour FROM flood_tweets_2013_2014 WHERE date_trunc('hour',
posted_time) AT TIME ZONE 'ICT' >= '2014-01-01' AND date_trunc('hour',
posted_time) AT TIME ZONE 'ICT' < '2014-03-02' GROUP BY hour) as tw LEFT
OUTER JOIN (SELECT count(objectid) as count, date_trunc('hour', end_date_time)
as hour FROM timeseries_flood_2013_2014 GROUP BY date_trunc('hour',
end_date_time) ) as fl ON tw.hour = fl.hour ORDER BY tw.hour
```

2. Codes for Python:

a). plot_flood_tweets_timeseries_sacatter :

```
import psycopg2 # database
import matplotlib.pyplot as plt # graphs
from numpy import *
#define transform function for different-hour-level
def data_transform(hour_level,orig_tweets,orig_floods,orig_time):
loop_nb=0
temp_ytw=0
temp_yfl=0
for j in range(0,len(orig_time)-1,1):
        if (x[j]-x[0+loop_nb*hour_level]).seconds<3600*hour_level:
        temp_ytw=temp_ytw+orig_tweets[j]
        temp_yfl=temp_yfl+orig_floods[j+1]
```

```
else:
        y_tweets_new.append(temp_ytw)
        y_floods_new.append(temp_yfl)
        x_new.append(x[j])
        loop_nb=loop_nb+1
        temp_ytw=orig_tweets[j]
        temp_yfl=orig_floods[j+1]
return;
#define function for plotting Flooded areas/tweets
def fl_tw(floods,tweets,time):
        tt=[]
        for i in range(0,len(floods),1):
                y=float16(floods[i])
                x=float16(tweets[i])
                tt.append(y/x)
plt.plot(time,tt,'*',label='Scatter  points')
plt.title('Scatter point of Flooded areas/Tweets')
plt.xlabel('Timeseries')
plt.ylabel('Flooded areas/Tweets')
plt.legend()
plt.savefig('Scatter point of Flooded areas_Tweets.png',dpi=600)
# database connection
conn = psycopg2.connect("dbname=datagrant user=Albert host=127.0.0.1")
# database cursor
```

```
cur = conn.cursor()

# get data

cur.execute("SELECT tw.count as tweet_count, coalesce(fl.count,0) as flood_count, tw.hour AT TIME ZONE 'ICT' FROM(SELECT count(id) as count, date_trunc('hour', posted_time) as hour FROM flood_tweets_2013_2014 WHERE date_trunc('hour', posted_time) AT TIME ZONE 'ICT' >= '2014-01-01' AND date_trunc('hour', posted_time) AT TIME ZONE 'ICT' < '2014-03-02' GROUP BY hour) as tw LEFT OUTER JOIN (SELECT count(objectid) as count, date_trunc('hour', end_date_time) as hour FROM timeseries_flood_2013_2014 GROUP BY date_trunc('hour', end_date_time) ) as fl ON tw.hour = fl.hour ORDER BY tw.hour")

data = cur.fetchall()

# organise data

y_tweets = []

y_floods = []

x = []

for i in data:

        y_tweets.append(i[0])

        y_floods.append(i[1])

        x.append(i[2])

"""new functions"""

# variables for new data INIL

y_tweets_new=[]

y_floods_new=[]

x_new=[]

#input the level (hours)

level=6

# use the data_transform function
```

```
data_transform(level,y_tweets,y_floods,x)


for kk in range(0,len(y_tweets_new)-2,1):
        print y_tweets_new[kk],y_floods_new[kk],x_new[kk]
#use the plotting flooded areas/tweets function
fl_tw(y_floods_new,y_tweets_new,x_new
# initialise plots using pylab
fig, host = plt.subplots()
fig.set_size_inches(12,8)
# add second y-axis
second = host.twinx()
# plot flood events
plot1=host.bar(x_new,y_floods_new,  color='#008eee', width=0.25, label='Flooded RW areas' )
host.set_ylabel('Number  of flooded "RW" neighbourhoods')
# plot Twitter activity
plot2=second.plot(x_new,y_tweets_new, color='green', label='Filtered Twitter activity')
second.set_ylabel('Number  of tweets')
# add plot elements
plt.title('Flooding  & Twitter activity daily - Jakarta, January 2014')
plt.legend()
fig.autofmt_xdate()


# save plot to file
```

```
fig.savefig('plot_flood_tweets_timeseries.png',dpi=1000)

fig.savefig('Scatter point of Flooded areas_Tweets.png',dpi=600)

# show plot

plt.show()

# close database

cur.close()

conn.close()
```

b). Spearman_rank_correlation:

```
from scipy.stats.stats import spearmanr

import psycopg2 # database

import matplotlib.pyplot as plt # graphs

from numpy import *


#define transform function for different-hour-level

def data_transform(hour_level,orig_tweets,orig_floods,orig_time):

    loop_nb=0

    temp_ytw=0

    temp_yfl=0


    for j in range(0,len(orig_time)-1,1):

        if (x[j]-x[0+loop_nb*hour_level]).seconds<3600*hour_level:

            temp_ytw=temp_ytw+orig_tweets[j]

            temp_yfl=temp_yfl+orig_floods[j+1]

        else:
```

```
            y_tweets_new.append(temp_ytw)
            y_floods_new.append(temp_yfl)
            x_new.append(x[j])
            loop_nb=loop_nb+1
            temp_ytw=orig_tweets[j]
            temp_yfl=orig_floods[j+1]
    return;

#define function for plotting Flooded areas/tweets
def fl_tw(floods,tweets,time):
    tt=[]
    for i in range(0,len(floods),1):
        y=float16(floods[i])
        x=float16(tweets[i])
        tt.append(y/x)

    plt.plot(time,tt,'*',label='Scatter  points')
    plt.title('Scatter point of Flooded areas/Tweets')
    plt.xlabel('Timeseries')
    plt.ylabel('Flooded areas/Tweets')
    plt.legend()
    plt.savefig('Scatter point of Flooded areas_Tweets.png',dpi=600)

# database connection
```

```
conn = psycopg2.connect("dbname=datagrant user=Albert host=127.0.0.1")


# database cursor

cur = conn.cursor()


# get data

cur.execute("SELECT tw.count as tweet_count, coalesce(fl.count,0) as flood_count, tw.hour AT TIME ZONE 'ICT'  FROM(SELECT count(id) as count, date_trunc('hour', posted_time) as hour FROM flood_tweets_2013_2014 WHERE date_trunc('hour', posted_time) AT TIME ZONE 'ICT' >= '2014-01-01' AND date_trunc('hour', posted_time) AT TIME ZONE 'ICT' < '2014-03-02' GROUP BY hour) as tw LEFT OUTER JOIN (SELECT count(objectid) as count, date_trunc('hour', end_date_time) as hour FROM timeseries_flood_2013_2014  GROUP BY date_trunc('hour', end_date_time) ) as fl  ON tw.hour = fl.hour ORDER BY tw.hour")

data = cur.fetchall()


# organise data

y_tweets = []

y_floods = []

x = []


for i in data:

   y_tweets.append(i[0])

   y_floods.append(i[1])

   x.append(i[2])


# variables for new data INIL
```

```
y_tweets_new=[]

y_floods_new=[]

x_new=[]

#input the level  (hours)

level=6


# use the data_transform function

data_transform(level,y_tweets,y_floods,x)


print y_tweets_new[20], y_floods_new[20],x[20]


print spearmanr(y_tweets_new,y_floods_new)


cur.close()

conn.close()
```

References

[1] S. E. Swanson, "GIS," *Hospital Librarianship,* vol. 1, pp. 83-89, 2008.

[2] D. J. Unwin, "GIS, spatial analysis and spatial statistics," *Progress in Human Geography,* vol. 20, pp. 540-551, 1996.

[3] L. Palen, K. Starbird, S. Viewg, and A. Hughes, "Twitter-based Information Distribution during the 2009 Red River Valley Flood Threat," *Bulletin of the American Society for Information Science and Tehcnology,* vol. 36, pp. 13-17, June/July 2010 2010.

[4] M. F. Goodchild, "Citizens as sensors: the world of volunteered geography," *GeoJournal,* vol. 69, pp. 211-221, 2007.

[5] T. Holderness, "Geosocial Intelligence," *IEEE Technology and Society,* vol. 33, pp. 17-18, 2014.

[6] S. J. Fusco, R. Abbas, K. K. Michael, and A. Aloudat, "Exploring the Social Implications of Location Based Social Networking: An inquiry into the perceived positive and negative impacts of using LBSN between friends," presented at the 9th IEEE International Conference on Mobile Business, Athens, Greece 2010.

[7] S. J. Fusco, R. Abbas, K. K. Michael, and A. Aloudat, "Location-Based Social Networking: Impact on Trust in Relationships," *Technology and Society Magazine, IEEE,* vol. 31, pp. 39-50, 2012.

[8] E. Akmalah and N. S. Grigg, "Jakarta flooding: systems study of socio-technical forces," *Water International,* vol. 36, pp. 733-747, 2011.

[9] R. Cybriwsky, "City profile: Jakarta," *Cities,* vol. 18, pp. 199-211, 2001.

[10] S. Donnan, "Jakarta to speed up tsunami warning system," *Financial Times,* p. 1, 2006.

[11] D. Gherghita-Mihaila, "How is Social Media Influencing the Way we Communicate?," *Acta Universitatis Danubius: Communicatio,* pp. 74-83, 2016.

[12] A. C. Andrew Crooks, Anthony Stefanidis and Jacek Radzikowski, "#Earthquake: Twitter as a Distributed Sensor System," *Transactions in GIS,* vol. 17, pp. 124-147, 2013.

[13] A. Aloudat, K. Michael, and J. Yan, "Location-Based Services in Emergency Management-from Government to Citizens: Global Case Studies," in *Recent Advances in Security Technology*, Australian Homeland Security Research Centre, Melbourne, 2007, pp. 190-201.

[14] J. Ewart and H. McLean, "Ducking for cover in the ‘blame game’: news framing of the ndings of two reports into the 2010–11 Queensland oods," *Disasters,* vol. 39, pp. 166-184, 2014.

[15] M. Haklay, A. Singleton, and C. Parker, "Web mapping 2.0: The neogeography of the GeoWeb," *Geography Compass,* vol. 2, pp. 2011-2049, 2008.

[16] D. M. Boyd and N. B. Ellison, "Social Network Sites: Definition, History, and Scholarship," *Journal of Computer-Mediated Communication,* vol. 13, pp. 210-230, 2007.

[17] F. Harvey, "To Volunteer or to Contribute Locational Information? Towards Truth in Labeling for Crowdsourced Geographic Information. In D. Sui, S. Elwood, & M. F. Goodchild (Eds.)," *Crowdsourcing Geographic Knowledge* pp. 31-42, 2011.

[18] P. Symeonidis, D. Ntempos, and Y. Manolopoulos, "Location-Based Social Networks. Recommender Systems for Location-based Social Networks," *Springer New York,* 2014.

[19] F. v. E. A. Horita, L. v. C. Degrossi, L. F. F. G. Assis, Alexander, and J. o. P. d. Albuquerque, "VGI in Disaster Management: systematic literature review," *Proceedings of the Nineteenth Americas Conference on Information Systems, Chicago, Illinois,* pp. 1-11, 2013.

[20] O. Roick and S. Heuser, "Location Based Social Networks - Definition, Current State of the Art and Research Agenda.," *Transactions in GIS,* vol. 17, pp. 763-784, 2013.

[21] T. Blaschke and C. Eisank, "How influential is Geographic Information Science," 2012.

[22] B. Kitchenham and S. Keele, "Guidelines for performing Systematic Literature Reviews in Software Engineering," *In EBSE Technical Report EBSE-2007-01,* 2007.

[23] P. Breretona, B. A. Kitchenham, D. Budgen, M. Turner, and M. Khalil, "Lessons from applying the systematic literature review process within the software engineering domain," *Journal of Systems and Software,* vol. 80, pp. 571-583, 2007.

[24] K. Watanabe, M. Ochi, M. Okab, and R. Onai, "Jasmine: a real-time local-event detection system based on geolocation information propagated to microblogs," *In CIKM '11 Proceedings of the 20th ACM international conference on Information and knowledge management. ACM New York, USA,* pp. 2541-2545, 2011.

[25] M. Sofean and M. Smith, "A Real-Time Architecture for Detection of Diseases using Social Networks: Design , Implementation and Evaluation," *Proceedings of the 23rd ACM conference on Hypertext and social media. ACM,* pp. 209-210, 2012.

[26] A. Veloso and F. Ferraz, "Dengue surveillance based on a computational model of spatio-temporal locality of Twitter," in *Proceedings of the 3rd International Web Science Conference. ACM.*, 2011.

[27] V. Lampos and N. Critianini, "Tracking the flu pandemic by monitoring the social web," Elba, 2010, p. 6.

[28] M. Gerais, "Traffic Observatory: a system to detect and locate traffic events and conditions using Twitter," in *Proceedings of the 5th International Workshop on Location-Based Social Networks*, ACM, 2012, pp. 5-11.

[29] T. Sakaki and Y. Matsuo, "Real-time Event Extraction for Driving Information from Social Sensors," in *2012 IEEE International Conference,* 2012, pp. 221-226.

[30] E. Adi and R. Kosala, "Harvesting Real Time Traffic Information from Twitter," *Procedia Engineering,* vol. 50, pp. 1-11, 2012.

[31] S. Wakamiya and R. Lee, "Crowd-sourced Urban Life Monitoring: Urban Area Characterization based Crowd Behavioral Patterns from Twitter Categories and Subject Descriptors," in *Proceedings of the 6th International Conference on Ubiquitous Information Management and Communication*, 2012, p. 26.

[32] L. Ferrari, A. Rosi, M. Mamei, and F. Zambonelli, "Extracting urban patterns from location-based social networks," in *LBSN '11 Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks*, New York, NY, USA, 2011, pp. 9-16.

[33] A. Sadilek, J. Krumm, and E. Horvitz, "Crowdphysics: Planned and Opportunistic Crowdsourcing for Physical Tasks," *SEA*, vol. 21, p. 2, 2013.

[34] G. Andrienko and N. Andrienko, "Thematic Patterns in Georeferenced Tweets through Space-Time Visual Analytics," *Computing in Science & Engineering*, vol. 15, pp. 72-82, 2013.

[35] S. Kinsella, V. Murdock, and N. O. Hare, "“ I ’ m Eating a Sandwich in Glasgow ”: Modeling Locations with Tweets," in *Proceedings of the 3rd international workshop on Search and mining user-generated contents*, 2011, pp. 61-68.

[36] B. H. e. al., "Tweets from Justin Bieber ’ s Heart: The Dynamics of the “ Location ” Field in User Profiles," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2011, pp. 237-246.

[37] B. Lee and B.-Y. Hwang, "A Study of the Correlation between the Spatial Attributes on Twitter," in *2012 IEEE 28th International Conference on Data Engineering Workshops*, 2012, pp. 337-340.

[38] R. Gonzalez and Y. Chen, "TweoLocator: A Non-Intrusive Geographical Locator System for Twitter," in *Proceedings of the 5th International Workshop on Location-Based Social Networks*, 2012, pp. 24-31.

[39] M. Pennacchiotti and A. Popescu, "A Machine Learning Approach to Twitter User Classification," in *ICWSM 12*, 2010, pp. 298-305.

[40] Y. Takhteyev, A. Gruzd, and B. Wellman, "Geography of Twitter networks. Social Networks," *34*, vol. 1, pp. 73-81, 2012.

[41] M. C. e. al., "Spatiotemporal social media analytics for abnormal event detection and examination using seasonal-trend decomposition," in *2012 IEEE Conference on Visual Analytics Science and Technology*, 2010, pp. 143-152.

[42] J. Weng, E. Lim, and J. Jiang, "Twitterrank: Finding Topic-Sensitive Influential Twitterers," in *Proceedings of the third ACM international conference on Web search and data mining*, 2010, pp. 261-270.

[43] B. Krishnamurthy and M. Arlitt, "A Few Chirps About Twitter," in *Proceedings of the first workshop on Online social networks*, 2006, pp. 19-24.

[44] A. Go, L. Huang, and R. Bhayani, "Sentiment Analysis of Twitter Data," *Entropy*, pp. 30-38, 2009.

[45] D. Quercia, L. Capra, and J. Crowcroft, "The Social World of Twitter: Topics , Geography , and Emotions," in *ICWSM*, 2012, pp. 298-305.

[46] H. J. Miller and M. F. Goodchild, "Data-driven geography," *GeoJournal*, 2014.

[47] Y. Levy and T. J. Ellis, "A Systems Approach to Conduct an Effective Literature Review in Support of Information Systems Research," *Informing Science*, p. 9, 2006.

[48] D. Ebdon, *Statistics in Geography: A Practical Approach - Revised with 17 Programs*, 2, illustrated, reprint, revised ed. New York, NY, USA: Wiley-Blackwell, 1985.